Contrasting paternal and maternal genetic histories of Thai and Lao populations

Wibhu Kutanan[1,2,*], Jatupol Kampuansai[3,4], Metawee Srikummool[5], Andrea Brunelli[6], Silvia Ghirotto[6], Leonardo Arias[2], Enrico Macholdt[2], Alexander Hübner[2], Roland Schröder[2], and Mark Stoneking[2,*]

[1]Department of Biology, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand

[2]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[3]Department of Biology, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand

[4]Center of Excellence in Bioresources for Agriculture, Industry and Medicine, Chiang Mai University, Chiang Mai, Thailand

[5]Department of Biochemistry, Faculty of Medical Science, Naresuan University, Phitsanulok, Thailand

[6]Department of Life Science and Biotechnology, University of Ferrara, Ferrara, Italy.

***Corresponding authors**:

Wibhu Kutanan (wibhu@kku.ac.th)

Mark Stoneking (stoneking@eva.mpg.de)

Abstract

The human demographic history of Mainland Southeast Asia (MSEA) has not been well-studied; in particular there have been very few sequence-based studies of variation in the male-specific portions of the Y chromosome (MSY). Here, we report new MSY sequences of ~2.3 mB from 914 males, and combine these with previous data for a total of 928 MSY sequences belonging to 59 populations from Thailand and Laos who speak languages belonging to three major MSEA families: Austroasiatic (AA), Tai-Kadai (TK) and Sino-Tibetan (ST). Among the 92 MSY haplogroups, two main MSY lineages (O1b1a1a* (O-M95*) and O2a* (O-M324*)) contribute substantially to the paternal genetic makeup of Thailand and Laos. We also analyse complete mtDNA genome sequences published previously from the same groups, and find contrasting pattern of male and female genetic variation and demographic expansions, especially for the hill tribes, Mon, and some major Thai groups. In particular, we detect an effect of post-marital residence pattern on genetic diversity in patrilocal vs. matrilocal groups. Additionally, both male and female demographic expansions were observed during the early Mesolithic (~10 kya), with two later major male-specific expansions during the Neolithic period (~4-5 kya) and the Bronze/Iron Age (~2.0-2.5 kya). These two later expansions are characteristic of the modern AA and TK groups, respectively, consistent with recent ancient DNA studies. We simulate MSY data based on three demographic models (continuous migration, demic diffusion and cultural diffusion) of major Thai groups and find different results from mtDNA simulations, supporting contrasting male and female genetic histories.

Keywords: Y chromosome, mtDNA, Austroasiatic, Tai-Kadai, Sino-Tibetan

Introduction

Thailand and Laos occupy a key location in the center of Mainland Southeast Asia (MSEA; Figure 1), which is undoubtedly one of the factors facilitating the extensive ethnolinguistic diversity there, as there are 68 recognized groups in Thailand and 82 groups in Laos, belonging to five language families (Simons and Fennig 2018). The prehistoric peopling of the area of present-day Thailand and Laos has been documented by several archaeological studies (Shoocongdej 2006; Demeter et al. 2012; Higham 2014; Higham 2017) and investigated further by recent ancient DNA studies (Lipson et al., 2018; McColl et al., 2018). The earliest presence of modern humans in SEA is dated to ~50 thousand years ago (kya) (Higham 2013; Bae et al. 2017), followed by Paleolithic migration to East Asia ~30 kya, inferred from genetic data (Yan et al. 2014; Hallast et al. 2015). There was also an expansion of Neolithic farmers and Bronze Age migrations from southern China to MSEA, which contributed to the present-day gene pool of modern MSEA people, e.g. Thais and Laotians (Higham 2014; Higham 2017; Lipson et al. 2018; McColl et al. 2018). Additional migrations during the historical period from neighboring countries (Penth 2000; Schliesinger 2000) have additionally enhanced ethnolinguistic diversity.

The census size for Thailand was ~68.41 million in 2017, and for Laos was ~6.76 million in 2016 (Simons and Fennig 2018). There are five linguistic families distributed in these two countries. While the Tai-Kadai (TK) language is widely spread in southern China and MSEA, it is concentrated in present-day Thailand and Laos as it is a major language spoken by Thais (90.5%) and Laotians (67.7%). Austroasiatic (AA) speakers are next most frequent, accounting for 4.0% in Thailand and 24.4% in Laos. In addition, this area is also inhabited by historical migrants who speak Sino-Tibetan (ST), Hmong-Mien (HM), and Austronesian (AN) languages (frequencies of 3.2%, 0.3% and 2% respectively in Thailand; 3.1%, 4.8% and 0% in Laos) (Simons and Fennig 2018).

It is generally thought that AA languages were brought to the Thai/Lao region by Neolithic farmers from southern China, while TK languages were brought by a later, Bronze Age migration, also from southern China (Bellwood 2018). The Neolithic expansion was ~2-3 kya before the expansion of TK languages; thus, the AA people were thought to be present before the TK expansion. The TK migration during the Bronze Age could have occurred via either demic diffusion (an expansion of TK people that brought both their genes and their language) or cultural diffusion (a language spread with minor movement of people). A genetic study on the origin of TK people supports a southern Chinese origin (Sun et al. 2013), while our previous studies of mtDNA genome sequences supports demic diffusion as the best explanation for the origin of the present-day Thai/Lao TK groups, although there is a strong signal of admixture between TK and AA groups in central Thailand (Kutanan et al. 2017; Kutanan et al. 2018b).

The male-specific portions of the Y chromosome (MSY) are paternally-inherited and exhibit lineages specific to populations/geographic regions, making the MSY an informative tool for reconstructing paternal genetic history and demographic change (Barbieri et al., 2014; Yan et al., 2014). However, to date there have been few MSY studies of MSEA and almost all of them employed short tandem repeats (Y-STRs) (Cai et al. 2011; Kutanan et al. 2011; Brunelli et al. 2018) which complicates comparison with mtDNA sequences due to their different mutational mechanism. In addition, those previous studies have also defined haplogroups by genotyping assays, which are thus biased in terms of the haplogroups detected and cannot uncover new sublineages.

We have previously carried out comprehensive studies of the maternal genetic history of the Thai/Lao region, based on 1,823 complete mtDNA genome sequences (Kutanan et al. 2017; Kutanan et al. 2018a; Kutanan et al. 2018b). In order to investigate the paternal genetic variation and demographic history, here, we investigate ~2.3 mB of MSY sequence in a subset of the above individuals, comprising 928 sequences from 59 populations. We compare and contrast the MSY and mtDNA results, and we also use demographic modeling to address the role of demic vs. cultural diffusion vs. admixture in the origins of the major TK groups in each Thai/Lao region.

Our MSY sequencing results provide new insights into the paternal genetic history of MSEA, and indicated contrasting paternal and maternal histories in this region.

Results

We generated 914 sequences of ~2.3 mB of the MSY, which combined with 14 published sequences brings the total to 928 MSY sequences belonging to 59 populations from Thailand and Laos (Figure 1; Table S5). There are 816 haplotypes defined by 8160 polymorphic sites, with mean coverages ranging from 4X to 109X (overall average coverage = 23X). Among the 928 MSY sequences, there are 92 specific haplogroups, belonging mostly to two main MSY lineages (O1b* and O2a*), that contribute substantially to the paternal genetic makeup of Thailand and Laos. There are several subclades of O1b*; the most frequent (50.54%) is O1b1a1a* or O-M95*, which occurs in almost half of the AA groups with a very high frequency (>70%), i.e. KH1-KH2, KA, BU, BL, SU, TN1-TN3, MA and LW3 (Figure 1: Table S1). The Correspondence Analysis (CA) (based on haplogroup frequencies) also supports the divergence of these AA speaking groups in agreement with the other results mentioned later, with many O1b* sublineages, e.g. O1b1a1a1b1a (O-B426) and O1b1a1a1a1a (O-F2758) (Figure S1). O2a* or O-M324* is the second most frequent haplogroup (25.86%) and has a relatively high frequency (>40%) in some AA and TK groups, and all ST speaking Karen. Additional minor non-SEA specific haplogroups were also observed, e.g. haplogroup N, found in the Lawa groups, and haplogroups R*, H*, and J*, which support associations between Indian and the Mon, and genetic connections between Mon and TK groups (Figure 1 and Figure S1). Further details on haplogroup distribution are provided in Table S1 and Supplementary Text.

Genetic diversity and structure

Generally, the AA populations show lower genetic diversity values than the TK and ST groups for the MSY, in agreement with the mtDNA results (Figure 2) (Mann–Whitney *U* tests between AA and TK for MSY: *h*: $Z = 3.37$, $P < 0.01$, MPD: $Z = 2.40$, $P < 0.05$, haplogroup diversity: $Z = 3.74$, $P < 0.01$ and for mtDNA: *h*: $Z = 4.33$, $P < 0.01$, MPD: $Z = 1.47$, $P > 0.05$, haplogroup diversity: $Z = 4.37$, $P < 0.01$). After the Maniq (MN), who have no MSY variation, and the Mlabri (MA), who have no mtDNA variation, the Htin (TN1), Lawa (LW3) and Bru (BU) show very low diversity values of MSY whereas the Htin (TN1-TN3), Khmer (KH2) and Seak (SK) show low mtDNA diversity (Figure 2). In contrast to the other AA groups, the Mon (MO1-MO7) show higher levels of both MSY and mtDNA diversity than other AA groups (Mann–Whitney *U* tests between AA and Mon for MSY: *h*: $Z = -3.33$, $P < 0.01$, MPD: $Z = -3.30$, $P < 0.01$, haplogroup diversity: $Z = -3.75$, $P < 0.01$ and for mtDNA: *h*: $Z = -1.94$, $P > 0.05$, MPD: $Z = -2.03$, $P < 0.05$, haplogroup diversity: $Z = -2.79$, $P < 0.01$). LW3 showed very low MSY haplogroup diversity and MPD values (Figure 2C), and a significantly low

Tajima's D value (Figure 2D), suggesting recent paternal expansion in this group, but the converse trend (rather high diversity) for mtDNA. Interestingly, a significantly negative Tajima's D value was observed more frequently in the TK than the AA groups for both the MSY and mtDNA (MSY, $P < 0.05$: 10/31 for TK vs. 6/24 for AA; mtDNA, $P < 0.05$: 20/31 for TK vs. 5/24 for AA) (Figure 2D), suggesting a stronger signal of recent population expansion in TK groups; no significant Tajima's D values were observed in any of the ST-speaking Karen groups. The Nyahkur (BO), who speak a Mon language, show the highest MPD value for the MSY, which might indicate paternal gene flow with other populations; this is supported by the BO having the highest number of shared MSY haplotypes (3 haplotypes) with other populations (Figure 3A). MO3 and MO4 have shared MSY haplotypes with the TK speaking groups (CT2, CT6 and YU1), reflecting their genetic connection. In the mtDNA, apart from the AA-speaking Palaung (PL), the Mon (MO2, MO3 and MO7) also share haplotypes with the central Thai (CT3 and CT6) and Shan (SH) (Figure 3A).

The Analysis of Molecular Variance (AMOVA) indicates that the variation among groups accounts for 11.20% of the total MSY genetic variance (Table 1). There is greater genetic heterogeneity among the AA groups (20.01%, $P < 0.01$ and 18.49%, $P < 0.01$ without MN, the hunter-gatherer group from southern Thailand) than among the TK (4.48%, $P < 0.01$) and ST-speaking Karen groups (2.29%, $P > 0.01$). For the AA groups with more than one population sampled, the greatest among-group variation by far was among the three Lawa populations (34.43%, $P < 0.01$), while the seven Mon populations showed very low (albeit still significant) among-group variation (3.92%, $P < 0.01$) (Figure S2). Very low among-group variation was also observed for the central Thai groups from central Thailand (1.47% $P > 0.01$), Khon Mueang groups from northern Thailand (-1.83%, $P > 0.01$), and Lao Isan groups from northeastern Thailand (1.84%, $P > 0.01$), indicating overall genetic homogeneity among these major TK speaking groups. In agreement with the MSY, larger mtDNA variation is observed in the AA groups (14.03%, $P < 0.01$) than the ST (6.51%, $P < 0.01$) and TK groups (4.33%, $P < 0.01$), but interestingly the largest among-group variation is not among the Lawa (7.78%, $P < 0.01$) but rather among the Htin populations (25.71%, $P < 0.01$). In contrast to the MSY, each of the TK groups with more than one population sampled showed significant among-group differences for mtDNA, especially the Khon Mueang (4.20%, $P < 0.01$) (Figure S2). In sum, we observed different patterns of MSY vs. mtDNA for the different language groups. The among-population variation within linguistic groups is larger for the MSY (20.01%, $P < 0.01$) than for mtDNA (14.03%, $P < 0.01$) for AA groups, but about the same for TK groups (4.48%, $P < 0.01$ for MSY and 4.33%, $P < 0.01$ for mtDNA), and the ST groups have larger among-population variation for mtDNA (6.51%, $P < 0.01$) than for the MSY (2.29%, $P < 0.01$) (Table 1; Figure S2). Thus, there are different patterns of MSY vs. mtDNA differentiation for these three language families.

Although there is more variation among groups defined by geographic location (2.38%, $P < 0.01$) than by language family (1.63%, $P < 0.01$) (Table 1) there is much more MSY variation among populations within the same group than among groups defined either by geographic or linguistic criteria. Moreover, when the

divergent MN population of hunter-gatherers from southern Thailand is removed from the analysis, then the among-group component is no longer significant for either geographic location or language family (-0.09%, $P > 0.01$ for geography; -0.01%, $P > 0.01$ for language), and the total variation among populations reduces to 10.54%. Thus, neither geography nor language family is a good predictor of the MSY genetic structure of Thai/Lao populations.

There are significant correlations between matrices of MSY genetic and geographic distance, estimated by Mantel tests, for all three types of geographic distances, i.e. great circle distance ($r = 0.3381$, $P < 0.01$), resistance distance ($r = 0.5418$, $P < 0.01$) and least-cost path distance ($r = 0.3912$, $P < 0.01$). However, the correlations are no longer significant when the MN group is removed from the analysis: great circle distance ($r = 0.0125$, $P > 0.05$), resistance distance ($r = -0.0446$, $P > 0.05$) and least-cost path distance ($r = 0.0139$, $P > 0.05$). In contrary, no significance was detected ($P > 0.05$) between matrices of mtDNA genetic distance and geographic distances with and without MN (great circle distance: $r = 0.0776$ and $r = -0.0323$), resistance distance ($r = 0.1433$ and $r = -0.1105$) and least-cost path distance ($r = 0.0997$ and $r = -0.0253$).

To identify and describe population clustering based on multivariate analysis, Discriminant Analysis of Principal Components (DAPC) was carried out. This analysis attempts to maximize among-groups genetic differentiation and minimize within-group genetic variation; the results showed considerable overlap among groups defined by either language family or geographic location in both MSY and mtDNA (Figure S3). In addition, the groupings by population and ethnicity of MSY data revealed the largest discrimination to be among some AA-speaking groups, i.e. all Lawa groups (LW1-LW3), Htin (TN1) and Blang (BL) whereas all Htin groups (TN1, TN2 and TN3), Mlabri (MA), TK-speaking Seak (SK) and ST-speaking Karen (KSK1, KSK2 and KPW) are differentiated from the others for mtDNA, emphasizing contrasting genetic pattern between MSY and mtDNA for Htin, Mlabri, Lawa, Blang, Seak and Karen.

In sum, all results indicate lower genetic diversity of the AA groups than the TK and ST groups, except the Mon and Nyahkur, who exhibit high genetic diversity. The AA groups also show greater genetic heterogeneity than the TK and ST groups.

Post-marital residence and genetic diversity

Although the influence of post-marital residence (patrilocal vs. matrilocal) has previously been studied in northern Thai hill tribes, these studies compared genetic variation between partial mtDNA sequences (hypervariable regions of the control region) and Y-STR loci (Oota et al. 2001; Besaggio et al. 20007). Here, we report the first comparison of mtDNA and MSY variation based on comparable sequence data. We studied four hill tribes (Karen, Htin, Lawa and Khmu) and the Palaung, another minority group in the mountainous area of northern Thailand but not officially recognized as a hill tribe. The Khmu (KA), Lawa (LW1, LW2 and LW3) and Palaung (PL) groups practice patrilocality (i.e., the wife moves to the residence of her husband after

marriage), whereas the Htin (TN1, TN2 and TN3) are matrilocal, as are the ST-speaking Karen (KSK1, KSK2, KPA and KPW). If residence pattern is influencing genetic variation, then lower within-population genetic diversity coupled with greater genetic heterogeneity among populations is expected for patrilocal groups than for matrilocal groups for the MSY, while the opposite pattern is expected for mtDNA (Oota et al. 2001). Mean values of *h,* MPD and haplogroup diversity in MSY are higher in matrilocal than patrilocal groups, not significantly different for *h* and MPD (Mann–Whitney *U* tests: *h*: $Z = 1.4616$, $P > 0.05$; MPD: $Z = 0.9744$, $P > 0.05$) but significantly different for haplogroup diversity (Mann–Whitney *U* tests: $Z = 2.1112$, $P < 0.05$) (Figure S4). For mtDNA, non-significant higher genetic diversity values for patrilocal than matrilocal groups are observed (Mann–Whitney *U* tests: *h*: $Z = -0.9744$, $P > 0.05$; MPD: $Z = -0.8120$, $P > 0.05$; haplogroup diversity: $z = -1.864$, $P > 0.05$) (Figure S4). Notably, TN1 and LW3 exhibit very low with-in population diversity for the MSY, e.g. MPD = 20.07 and 23.07, compared to the average MPD (121.11), whereas TN1 and TN2 (20.69 and 26.14) show lower MPD than average (35.09) for mtDNA. For genetic differences between-populations, the patrilocal Khmu, Lawa and Palaung have significantly higher genetic differentiation for the MSY than for mtDNA (average $\Phi_{st} = 0.3109$ for MSY and 0.0774 for mtDNA) (Mann–Whitney *U* tests: $Z = 3.5907$, $P < 0.01$) whereas the matrilocal groups (Htin and Karen) also show higher average Φ_{st} for MSY (0.1859) than for mtDNA (0.1553), but these are not significantly different (Mann–Whitney *U* tests: $Z = 0.3270$, $P > 0.05$). Contrasting genetic differences for the MSY vs. mtDNA of Lawa, Htin and Karen are clearly seen in the MDS and DAPC plots (Figure 4A and 4B; Figure S3). Much stronger contrasting between-group variation is seen in the AMOVA results (Lawa: 34.43% for MSY and 7.78% for mtDNA; Htin: 11.53% for MSY and 25.71% for mtDNA; Karen: 2.29% for MSY and 6.51% for mtDNA (Table 1; Figure S2).

However, in general, the AA-speaking groups, whether identified as hill-tribes or as other minorities, are patrilocal groups. The AMOVA result indicates that the variation among AA populations is higher in MSY (20.01%) than mtDNA (14.03%), in accordance with expectations if residence pattern is influencing genetic variation. Conversely, the TK populations, where neither patrilocal nor matrilocal residence is preferred, exhibit similar among-population variances for the MSY (4.48%) and mtDNA (4.33%) (Table 1; Figure S2). Overall, there does seem to be some impact of post-marital residence on the patterns of genetic diversity.

Genetic relatedness among populations

The genetic distance and MDS analyses based on MSY and mtDNA indicate that the MN and MA are highly diverged from the other populations for the MSY and mtDNA, respectively (Figure 3B and S5). The MA and MN also show large differences from the other populations in the heat plots of Φ_{st} values (Figure S5). However, in general both MSY and mtDNA results show relatively larger genetic heterogeneity of the AA groups vs. genetic homogeneity of the TK and ST groups (Figure 3B). The Mantel test of Φ_{st} values showed a significant correlation between the MSY and mtDNA Φ_{st} matrices ($r = 0.4506$, $P < 0.01$). After excluding these

MA and MN as outliers, the MDS for the MSY showed that almost all AA speaking groups are located along the edges of the plot, while most of the TK groups cluster in the center of the plot (Figure 4A), further supporting genetic heterogeneity of the AA and homogeneity of the TK populations. Interestingly, the SEA-specific O-M95* and O-M234* haplogroups (with several sublineages) differentiate the studied populations into at least two main paternal sources, and the frequencies of these two haplogroups correspond to the major differentiation in the MDS plot (Figure 4A). O-M95* is at high frequency in the populations on the left of the plot and gradually decreases to very low frequency in the populations on the right side in the first dimension, whereas the O-M324* frequency runs opposite to the O-M95* cline: O-M324* is at higher frequency in populations located on the right of the plot and decreases in frequency toward the left side (Figure 4A). The MDS plot and heat plot of MSY also indicates some Mon groups (MO1, MO3, MO5 and MO6) are close to the cluster of TK groups in the center of the plot (Figure 4A and 4C), indicating a close genetic relationship. In addition, non-SEA haplogroups lineages, e.g. R*, H*, and J*, provide more support for genetic connections between Mon and Central Thais.

For the MDS based on mtDNA (Figure 4B), the Mon generally showed genetic affinities with the TK groups in the center of the plot, with the exception of MO1, MO5 and MO6, which differ from the other Mon groups, as can be also seen in the MDS plot and heat plot (Figure 4B and 4D). Overall, we observe more genetic heterogeneity of the AA groups than the other linguistic groups and there are contrasting patterns of genetic relationships for the MSY vs. mtDNA.

Genetic relatedness between Thai/Lao and other Asian populations

The MDS based on the MSY Φ_{st} matrix of 73 populations from across Asia revealed that, in general, population clustering largely reflects linguistic affiliation (Figure 5), with some exceptions. In the first and second dimension, the AA populations are the most diversified, with the PL and MN appearing as outliers. There is one cluster of AA populations on the left, which also includes one TK group (BT2); the other AA populations are scattered along the main axis of the plot. Some Mon groups (MO2, MO4 and MO7) are relatively close to Indian and ISEA populations, indicating potential connections. Two central Thai groups (CT4 and CT7) are also relatively close to the Indian populations. The ST populations (Karen, Han Chinese and Burmese) are rather close. The ISEA and Papuan populations are in closer proximity to South Asian populations (Indian, Bengali and Punjabi). Generally, the haplogroup profile indicates genetic affinities between the Mon and South/Central Asian groups, which is consistent with the MDS plots (Figure 5) and results from previous mtDNA haplogroup analyses (Kutanan et al. 2017; Kutanan et al. 2018b).

The expansion of male lineages

The Bayesian Skyline Plots (BSP) of effective population size change (N_e) over time in each group reveal overall 5 different trends (Figure 6). The most common trend, found in Mon, Khmer, Htin, Central Thai and Black Tai, showed N_e increasing gradually or remaining constant during 40-60 kya until a decline ~5-7 kya,

followed by rapid growth ~5 kya and then a decrease ~2.0-2.5 kya. The other trends differ from the first trend as follows: no population reduction ~2.0-2.5 kya but population size either increases (Khon Mueang and Yuan) or remains stable (Lao Isan and Laotian); the Lue and Phuan show two increases in N_e, at about ~5 kya and ~10 kya; the Lawa show a stable population size since ~30 kya and then a decline during the last 2 kya with a sudden increase ~1 kya; and the Karen differ only slightly from the common trend with a population increase ~1 kya.

By contrast, the BSP based on mtDNA sequences for each ethnicity show three common trends (Figure 6). The first trend is an increase in N_e during 40-50 kya, followed by stability and then decrease ~2 kya, which was observed in Mon, Htin, Lawa, Khmer, Yuan, Phuan and Lue. The second pattern, shown by the Khon Mueang, is an increase in N_e ~40-50 kya, followed by stability and then increase again ~10 kya, followed by a decline ~2 kya. The Central Thai, Lao Isan and Laotian show the third trend, in which population increases occur ~40-50 and ~10 kya. In general, the BSP plot by ethnicity indicated lower effective population sizes for the MSY than for mtDNA (Figure 6).

We also plotted the BSP of several Asian populations from published MSY data (Karmin et al. 2015; Poznik et al. 2016) (Figure 7). Almost all of the MSEA and East Asian populations, i.e. Kinh, northern Han, southern Han and Japanese show a pronounced increase of the MSY N_e during ~4-6 kya, except the Xishuangbanna Dai, in which there is an increase ~2 kya. Around 5 kya, the Japanese show a decrease in N_e before a sudden increase, suggesting a bottleneck prior to demographic expansion. Interestingly, the ISEA population shows a large increase in N_e ~35-40 kya and a smaller increase ~2.5-3 kya. The South Asian populations, i.e. Bengali, Punjabi and Indian, also show two pulses of population increase at about the same times. The Punjabi also show an additional small increase in N_e change during ~12 kya.

The BSP by each major MSY haplogroup show four pulses of paternal N_e increases, at ~9-11 kya, ~5 kya, ~2.0-2.5 kya and ~1.0 kya (Figure 8), in agreement with the plot by ethnicity. The early Holocene N_e increment is obviously noticed in O2a1c* and O2a2a*, whereas the N_e growth ~5 kya is observed in O1b1a1a1b* and R*. Haplogroup O1a*, C* and D* show expansions in N_e ~2.0-2.5 kya and haplogroup N* shows a recent expansion ~1.0 kya. In addition, there are two expansion times for O1b1a1a1a* and O2a2b* (~5 and ~2 kya).

Demographic models

Previously, we used mtDNA genome sequences and demographic modeling to test different hypotheses about the origins of TK groups. Specifically, we tested whether different TK groups were primarily related to local AA groups (reflecting cultural diffusion, i.e. an AA group switching to a TK language), to a TK group from southern China (reflecting demic diffusion, i.e. spread of TK languages via migration from southern China), or were related to both (reflecting admixture between an incoming TK group from southern China and a local AA group). We found that the Khon Mueang (from northern Thailand), Lao Isan (from northeastern Thailand) and Laotian most likely originated via demic diffusion from southern China without substantial gene flow from

AA groups (Kutanan et al. 2017). However, for the central Thai, the most likely scenario was demic diffusion with a very low level of gene flow between central Thai and Mon groups (Kutanan et al. 2018b). Here we use the same approach to test three demographic scenarios concerning the paternal origins of these major Thai groups (Figure S6).

For the Khon Mueang (KM) people (Test 1), the highest posterior probability (0.80) and rather highly selected classification trees (0.58) were found for the demic diffusion model (Table S2). By contrast, the cultural diffusion model is the most likely scenario for the Lao and central Thai groups. Both the combined Laotian (LA) and Lao Isan (IS) datasets (Test 2) and the separate LA dataset (Test 3) weakly support the cultural diffusion model (for Test 2; posterior probability = 0.56 and selected classification tree = 0.37 and for Test 3; posterior probability = 0.56 and selected classification tree = 0.39). The IS dataset (Test 4) supports cultural diffusion (with the present-day IS groups descended from local Khmer (KH) with the highest posterior probability (0.71) and classification trees selected slightly more often than for the other models (0.49). For Test 5 (the Central Thai (CT) dataset), the cultural diffusion model had the highest posterior probability at 0.58 and was selected slightly more often among the classification trees (0.50) than the other models. However, a Principal Component Analysis (PCA) plot shows that based on the first two PCs the observed data fall within the distributions simulated under the three models in only Test 4, whereas the other datasets fall within the simulated distributions for PCs 3 and 4, suggesting that there is low efficiency to reconstruct the variability of the observed data (Figure S7). The parameter estimation for the best performing models in all five tests was able to obtain point estimates for each of the simulated effective population sizes. However, the posterior distributions were generally flat (Table S3: Figure S8). We also calculated the MSY Φ_{st} and corrected pairwise difference among groups of populations used in ABC tests to estimate their genetic relationships (Table S4). The KM are closer to the Dai than the local AA group (Test 1), the ethnic Lao and Laotian showed similar genetic differences to both Dai and AA groups (Test 2 and Test 3), whereas the CT groups (Test 5) have closer genetic relationships to the local AA group than to Dai. In contrast, mtDNA Φ_{st} and corrected pairwise difference revealed that the KM and ethnic Lao are closer to the Dai than local AA while the CT exhibited somewhat similar genetic distances to both Dai and AA. Overall, the simulations based on MSY sequences, compared with previous mtDNA simulation together with tests of genetic difference by Φ_{st} and corrected pairwise differences, suggest different demographic histories for males and females in the region.

Discussion

In order to gain more insights into MSEA genetic history, we here investigate the paternal genetic variation and structure by sequencing ~2.3 mB of the MSY from representative ethnolinguistic groups from Thailand and Laos. In sum, most of the studied populations exhibit two major MSY haplogroups, O-M324* and

O-M95* in different proportions, indicating two major paternal sources. O-M324* was widely spread in the TK groups, while O-M95* is predominant in the AA groups. However, some TK populations (BT2 and IS3) and some AA populations (PL, BO and MO4) exhibited the opposite pattern (Figure 1; Table S1). We also compared patterns of MSY variation with mtDNA in the same set of populations and found some similar results, e.g. overall lower genetic diversity and greater heterogeneity of AA groups than of TK and ST groups, large differences between the Mon and the other AA groups, and genetic connections between the Mon and central Thai (Figure 2-4). However, in many respects the patterns of MSY and mtDNA variation are different, suggesting contrasting paternal and maternal genetic histories.

Factors influencing contrasting genetic variation in the hill tribes

Although the genetic variation of the studied populations appears to be influenced by both linguistic and geographic factors, when the very diverged Maniq group is removed, neither language nor geography impacts genetic variation, indicating that these two factors are not important in the broad view (Table 1). Other factors, i.e. cultural practices, admixture, and genetic drift, seem to be more influential. In the case of the hill tribes, post-marital residence and preservation of their identity are putatively influential factors. In Thailand, there are nine ethnic groups which are officially identified as hill tribes, i.e. the AA-speaking Lawa, Htin and Khmu, the HM-speaking Hmong and IuMien and the ST-speaking Karen, Lahu, Akha and Lisu. The Akha, Lisu, Hmong, IuMien, Lawa and Khmu practice patrilocality while the Lahu, Karen and Htin are matrilocal. If post-marital residence is influencing patterns of genetic variation, then the expectation is for larger between-group differences and smaller within-group diversity for patrilocal groups for the MSY, and the same trends for matrilocal groups for mtDNA (Oota et al. 2001). The first comparative study of mtDNA and MSY variation in patrilocal vs. matrilocal groups was carried out in northern Thai hill tribes, because they include both patrilocal and matrilocal groups within a small geographic scale, and found a strong impact of post-marital residence on the mtDNA and MSY variation (Oota et al. 2001). However, previous studies compared genetic variation between partial mtDNA sequences and Y-STRs (Oota et al. 2001; Besaggio et al. 20007); here we report the first comparison of mtDNA and MSY variation based on comparable sequence data.

We analyzed the Khmu, Palaung and Lawa groups, who practice patrilocality, whereas the Htin are matrilocal, similar to the ST-speaking Karen. The within-population genetic diversity values is in agreement with expectation, i.e. greater diversity of matrilocal than patrilocal groups for MSY and the opposite trend in mtDNA (Figure S4). Moreover, genetic differentiation between populations is correlated with post-marital residence. However, in many cases the differences between patrilocal and matrilocal groups are not significant, indicating that other factors are also having an effect. One factor in particular that could influence the within-population genetic diversity and between-population differentiation is geographic isolation, which enhances genetic drift, thereby lowering within-population genetic diversity and increasing between-population

differentiation. This could explain the very low internal diversity of some AA groups (Figure 2A-2C) and high differentiation from other groups, e.g. some groups of Htin (TN1) and Lawa (Figure 4A and 4B; Figure S3) that live in mountainous, isolated parts of northern Thailand. The Lawa furthermore favor intra-marriage (Nahhas 2007), which would also reduce genetic variation in this group.

Moreover, while the expected difference between patrilocal and matrilocal groups holds in some regions (Oota et al. 2001; Besaggio et al. 2007), in other regions patterns of mtDNA and MSY variation do not conform to expectations (Kumar et al. 2006; Arias et al. 2018), which is to be expected if other factors are also influencing patterns of genetic variation (Wilkins and Marlowe 2006).

Genetic variation and origin of the Mon

The Mon groups showed genetic differences from other AA populations but closer relatedness to the TK populations, especially the central Thai, in both MSY and mtDNA (Figure 2A, 2B, 3A and 3B). Our previous simulation results, based on mtDNA, also supported admixture among the Mon and central Thai groups (Kutanan et al. 2018b). In addition, some Mon groups (MSY: MO3, MO5, MO6 and mtDNA: MO2, MO3 and MO4) exhibit genetic affinities with the Karen (Figure 3B), reflecting genetic heterogeneity and contrasting genetic patterns between MSY and mtDNA. Admixture might be an important factor influencing the genetic structure of the lowland AA-speaking Mon. Archaeological evidence indicates that the Dvaravati civilization of the Mon was centered in present-day central Thailand and southern Myanmar, and had expanded to a large part of mainland Southeast Asia during the 6th to 7th century A.D. (Diffloth 1984; Guillou 1999; Saraya 1999). After the intensification of Thai and Burmese kingdoms, the Mon in Myanmar were conquered by the Burmese during the 18th century A.D.; the ethnic Mon in Myanmar are currently concentrated in the Mon and Karen States (Pon Nya 2001). In Thailand, the present-day Mon are distributed in central Thailand and surrounding areas, with some groups living in the North and the Northeast. However, they are not considered to be the descendants of the ancient Mon Dvaravati civilization in Thailand, but rather political refugees that fled from Myanmar to Thailand during the 16th to 19th centuries A.D. (Ocharoen 1998). However, based on linguistic evidence, the remnants of the Dvaravati Mon population are now considered a distinct ethnic group known as the Nyahkur (BO) whose communities are restrict found in hilly areas along the border between central and northeastern Thailand (Diffloth 1984). In contrary to linguistic evidence, the Nyahkur has no shared haplotype or related to any specific Mon groups, indicating their genetic differences. However, Nyahkur show genetic sharing in both MSY and mtDNA with the Khmer groups (Figure 3A) which reflects their previous connection. In addition, the high frequency of MSY haplogroup O2a* and C* (Figure 1), close genetic relationship to many TK and ST speaking groups (Figure 3B) and highest MPD value for MSY (Figure 2C) indicated later extensive gene flow, promoting the paternal difference of Nyahkur from the Mon and also other AA groups.

Previous genetic studies of G6PD mutations reported a high prevalence of the Mahidol type G6PD deficiency in the Mon, Burmese, and Karen, different from Thai, Laotian and Khmer groups exhibiting the Vientiane-type G6PD mutation (Iwai et al. 2001; Matsuoka et al. 2005; Nuchprayoon et al. 2008). Thus both our results and previous studies indicate a close genetic relationship among Mon, Burmese, and Karen in Myanmar, suggesting a common origin or extensive gene flow. Our previous mtDNA study also revealed genetic relations between some Mon groups (MO1 and MO5) and Burmese, with both of them close to some Indian populations, whereas other Mon groups are closer to the Karen groups (MO2, MO3, MO4) (see details in Kutanan et al. 2018b). In general, genetic mixing among Mon, Karen and Myanmar might have happened before the arrival of the Mon to Thailand, whereas mixing among the Mon and central Thai would have occurred after the arrival of the Mon. However, MSY data for the Burmese are limited, and further MSY studies of populations from Myanmar are needed to confirm this scenario.

A connection between Indian groups and the Mon is suggested by South/Central Asian MSY lineages in the Mon, e.g. R*, H*, J*, L* and Q* (Figure 1), consistent with some mtDNA lineages, e.g. W3a1b, M6a1a, M30, M40a1, M45a, and I1b (Kutanan et al. 2017; Kutanan et al. 2018b). Thus, both mtDNA and the MSY indicated contact between the ancestors of the Mon and Indian. Archaeological evidence also suggests Indian influences, e.g. the symbolism on the Dvaravati coin which indicates the importance of royalty, and includes several motifs associated with Indian precedents of the 1st-4th century A.D (Higham and Thosarat 2012).

Demographic changes

Demographic expansion of Thai/Lao populations are noticeably detected in both paternal and maternal lineages at the beginning of the Holocene, ~10 kya (Figure 6). In this period, increasing and more stable temperatures might have facilitated population expansion (Wen et al. 2016). The male N_e increase during the Holocene is primarily driven by the O2a2a* and O2a1c* lineages (Figure 8). The Holocene expansion might thus be related to an expansion of HM paternal lineages, as O2* (O-M122*) is thought to have arisen at the beginning of the Holocene near Tibet (van Driem 2017). According to this hypothesis, the bearers of this haplogroup became the progenitors of the "Yangtzean" or Hmong-Mien paternal lineages, and contributed this lineage to the ancient AA who carried O1b1a1a* or O-M95* by sharing of knowledge about rice agriculture. However, further sequencing of MSY lineages belonging to the HM populations are needed to verify this hypothesis.

During the Neolithic period, other significant expansions are observed in almost all ethnicities and many MSY haplogroups, i.e. O1b1a1a1b*, O1b1a1a1a* and R* (Figure 8). Previously, it was suggested that the demographic expansion pattern in the Neolithic in SEA shows strong expansion dynamics, different characteristics than the Paleolithic expansion, and sex-specific expansion patterns, with earlier expansions in female than in male lineages. (Wen et al. 2016). The expansion signals in our results coincide with the beginning of the SEA Neolithic ~5 to 4.5 kya, during which farming expanded from China to SEA (Bellwood 2018). The farming technology for food production could support a higher population density than hunting-gathering, as

agriculture could produce a more steady food supply, and males could avoid hunting dangerous animals; thus, effective population size would increase (Jobling et al., 2004; Yan et al., 2014). The farmer expansion ~4 kya was probably related to ancestral AA speaking hill tribes with predominantly O-M95* lineages that knew rice agriculture (van Driem 2017; Lipson et al. 2018; McColl et al. 2018). However, the movement of Neolithic groups from southern China to MSEA probably involved not only AA groups but also TK groups (Bellwood 2018). In our study, a Neolithic expansion signal was observed for the MSY in all studied groups, indicating a large demographic expansion and probable admixture among the ancestors of indigenous southern Chinese groups during the Neolithic period. Haplogroup R1a was previously suggested to show a similar expansion, with paternal population growth during ~6.5 to 4 kya observed globally (Poznik et al. 2016; Wang et al. 2016).

In addition, we found another significant expansion during the Bronze age ~2 kya that involves TK speaking populations, reflected by some haplogroups prevalent in the TK, e.g. O1a* (Figure 8). This TK related expansion is consistent with the strong expansion detected in the BSP of Xishuangbanna Dai (Figure 7) and corresponds with the results of a recent ancient DNA study (McColl et al., 2018). The southward expansion of the indigenous southern Chinese TK speakers to MSEA was probably driven by the Han Chinese expansion from the Yellow River basin to southern China during the Qin dynasty, starting ~2.5 kya (Belwood 2018). The migration and expansion of prehistoric TK groups during the Bronze Age has had a profound influence on the modern Thais and Laotians in term of languages and genes. Nowadays TK languages are mostly concentrated in present-day Thailand and Laos, and the relatively high level of TK genetic homogeneity might be also driven by this recent expansion.

Our previous mtDNA modeling to explore the migration and expansion of prehistoric TK groups during the Bronze Age supported the spread of TK languages via demic diffusion and admixture (Kutanan et al. 2017; Kutanan et al. 2018b). Here, a similar modeling approach for the MSY data found weak support for cultural diffusion of TK languages. Although we built the model based on historical sources, the models did not generate the observed variation (Figure S6 and S7), indicating that the analysed models do not correspond to the real paternal population history. A possible reason for this striking difference between maternal and paternal histories might be warfare. Historically, many areas of Thailand saw frequent warfare involving various TK groups ~200-500 ya (Penth 2000). As a result, forced migrations were imposed upon the losing side and men were taken captive more often than women because men could be used to strengthen the victors' armies. This could result in a different history for the TK male vs. female population. More complex demographic models could therefore more accurately capture the paternal history of Thai/Lao populations.

It may be that the MSY sequences do not harbor enough information to distinguish among the different demographic scenarios. However, comparison of genetic differences (Φ_{st} and corrected pairwise differences) among the groups used in the simulations does support a real contrast in the maternal vs. paternal histories for the major TK groups in each region, and also finds genetic heterogeneity among these major groups. The northern Thai people showed closer genetic relationship with the Dai than AA groups in both mtDNA and MSY,

supporting the demic diffusion model, whereas the ethnic Lao are closer to Dai for mtDNA but for MSY they are related to both Dai and AA rather equally, suggesting demic diffusion for the maternal history and admixture for the paternal history. The central Thai MSY sequences could be of AA origin because they are genetically more similar to the AA groups than the Dai, supporting cultural diffusion, but for mtDNA they are related to both Dai and AA rather equally, supporting admixture in central Thailand as found previously (Kutanan et al. 2018b). Overall, these results suggest that the demographic history of Khon Mueang, ethnic Lao and central Thais are different, possibly reflecting either different migration routes or different small TK groups that expanded from China (Higham and Thosarat 2012). In addition, different patterns of admixture for males vs. females could have occurred in ethnic Lao and central Thais. Archaeological and historical evidence indicate that prior to the TK migration, there were existing rich civilizations in the area, e.g. the Dvaravati of the Mon and Chenla of the old Khmer. With the arrival of TK groups, the Mon people were incorporated by intermarriage into Tai society and adopted the increasing dominant Thai language as their own (Higham and Thosarat 2012). Our results suggest that there was variation in the pattern of cultural diffusion/admixture involving males vs. females in different groups in the area of northeastern and central Thailand and Laos.

Finally, another more recent expansion signal was detected in the northern Thai AA-speaking Lawa, involving haplogroups O2a2b* and N* (Figure 6 and 8). Historical evidence indicates that after the arrival of the TK groups in northern Thailand, the native Lawa groups were fragmented and moved to the mountains (Penth 2000), resulting in cultural and geographical isolation. In support of this model of isolation and drift, we note that the most negative Tajima's D value is observed in the LW3 group, which suggests population expansion after a bottleneck (Figure 2D).

Conclusion

Several factors, e.g. cultural practices, gene flow, genetic drift and geography have influenced the genetic variation and genetic structure of present-day Thai/Lao populations. Here we compared high-resolution mtDNA and MSY sequences and found contrasts in the maternal and paternal genetic history of various Thai/Lao groups, in particular the hill tribes and the AA and ST speaking groups, as well as significant genetic heterogeneity among samples from the same ethnolinguistic group from different locations (Figure 1 and 4). Finally, this new MSY study from Thai/Lao males provides more insight into the past demographic history in the paternal line and, along with our previous mtDNA studies, is generally in agreement with recent ancient DNA studies in SEA that indicate two demographic expansions from southern China to MSEA, with the first involving the ancestors of AA groups and the second involving TK groups (Lipson et al. 2018; McColl et al. 2018). Overall, the contrasting results for the maternal vs. paternal history of some Thai/Lao groups supports the importance of detailed studies of uniparental markers, as such contrasts would not have been revealed by studying autosomal markers in just a few Thai/Lao groups. Additional ancient DNA studies, coupled with more

detailed genome-wide data from present-day populations, will provide a complete reconstruction of the genetic history of this region.

Material and methods

Studied populations

Genomic DNA was extracted from blood, buccal swab or saliva of 914 males belonging to 57 populations that were classified into 26 ethnolinguistic groups, as described previously (Kutanan et al. 2017; Kutanan et al. 2018a) (Figure 1; Table S5). Ethical approval for this study was provided by Khon Kaen University, Naruesuan University, and the Ethics Commission of the University of Leipzig Medical Faculty.

MSY sequences

We prepared genomic libraries for each sample using a double index scheme (Kircher et al. 2012) and enriched the libraries for ~2.34 mB of the MSY via in-solution hybridization-capture using a previously-designed probe set (Kutanan et al. 2018b) and the Agilent Sure Select system (Agilent, CA, USA); further details on the probe design are provided in Table S6. Sequencing was carried out on the Illumina HiSeq 2500 platform with paired-end reads of 125 bp length. Standard Illumina base-calling was performed using Bustard. Illumina adapters were trimmed and completely overlapping paired sequences were merged using leeHOM (Renaud et al. 2014). De-multiplexing of the pooled sequencing data was done by deML (Renaud et al. 2015). The alignment and post-processing pipeline of the sequencing data was described previously (Kutanan et al. 2018b).

Statistical analysis

Genetic diversity and structure

We combined the 914 newly-generated sequences together with 14 published sequences (Kutanan et al. 2018b) belong to two hunter-gatherer populations from Thailand: Mlabri and Maniq (Table S5). This study thus includes 928 MSY sequences from 59 populations and 28 ethnolinguistic groups of Thailand and Laos. To compare with the MSY data, we selected 1,434 mtDNA sequences from the same populations from our previous studies (Kutanan et al. 2017; Kutanan et al. 2018a; Kutanan et al. 2018b) (Table S5). We used Arlequin 3.5.1.3 (Excoffier and Lischer 2010) for the following analyses: summary statistics of genetic diversity within populations, the matrix of genetic distances (Φ_{st}), Analyses of Molecular Variance (AMOVA), and Mantel tests of the correlation between genetic and geographic distances.

Genetic relationships

To investigate the paternal relatedness between populations, we performed a discriminant analysis of principal components (DAPC) (Jombart et al. 2010). We grouped our samples based on population sampled, geographic location and ethnicity (Table S5) before running the analysis for 100,000 iterations using *adegenet* 1.3-1 (Jombart et al. 2008).

A correspondence analysis (CA) based on MSY haplogroup counts was performed using STATISTICA 13.0 (StatSoft, Inc., USA). Haplogroup assignment was performed by yHaplo (Poznik 2016). The R package (R Development Core Team) was used to carry out a nonparametric multidimensional scaling (MDS) analysis (based on Φ_{st} values of MSY and mtDNA), the MDS heat plot with 5 dimensions, showing per-dimension standardized values between 0 and 1, and heat plots of the Φ_{st} distance matrix and the matrix of shared haplotypes.

To get a broad picture of population relationships in Asia, we included 552 MSY sequences from Asian groups for comparison We downloaded the published Y chromosome sequencing data from the SGDP data set (https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/Y-bams/Y.tar) (Mallick et al. 2016), the 1000 Genomes Project (1000 Genomes Project Consortium et al. 2015) and the study of Poznik et al. (2016). We merged and processed all sequencing data through the same pipeline as the samples in our study (Kutanan et al. 2018b). The resulting variant file was merged with data from previous study (Karmin et al. 2015; http://evolbio.ut.ee/chrY/) using Heffalump v0.2 (https://bitbucket.org/ustenzel/heffalump). We subset the variant file to sites that were overlapping the regions present on our capture bait and to samples that had a major haplogroup that was also present in our data set. These samples were combined with our samples; we then removed variant sites for which < 25% of the samples had genotype information, and samples that had > 25% of all sites with missing genotype information. The resulting data set provides 16,684 variable sites, which was imputed using BEAGLE v4.1 (Browning and Browning 2016). Additional details on these populations are provided in Table S7.

Bayesian Skyline Plots

Based on Bayesian Markov Chain Monte Carlo (MCMC) analyses, BEAST 1.8.4 was used to construct Bayesian Skyline Plots (BSP) by ethnicity and by haplogroup (Drummond et al. 2012). To avoid a false detection of bottlenecks stemming from the sample collection procedure (Heller et al., 2013), we pooled all populations within the same ethnicity and ran jModel test 2.1.7 (Darriba et al. 2012) to select the most suitable model for each run during the creation of the input file for BEAST via BEAUTi v1.8.2. We used an MSY mutation rate of 8.71×10^{-10} substitutions/bp/year (Helgason et al. 2015), and the BEAST input files were modified by an in-house script to add in the invariant sites found in our dataset. Both strict and log normal relaxed clock models were run for each ethnicity and haplogroup, with marginal likelihood estimation (MLE) (Baele et al. 2012; Baele et al. 2013). After each BEAST run, the Bayes factor was computed from the log marginal likelihood of both

models to choose the best-fitting BSP plot. Tracer 1.5.0 was used to check the results. We also performed the BSP of compared populations, i.e. Dai, Kinh, Southern Han, Northern Han and Japanese from published MSY sequences (Poznik et al. 2016). The BSPs by ethnicity based on mtDNA genomes were carried out in a previous study (Kutanan et al. 2018a).

Approximate Bayesian Computation

In order to investigate the paternal origin of TK groups in Thailand/Laos and their local histories, we employed 5 datasets (encompassing northern Thailand, central Thailand, and northeastern Thailand and Laos) and compared 3 competing scenarios: demic diffusion (i.e., a migration of people from southern China, who are then the ancestors of present-day Thai/Lao TK people); cultural diffusion (i.e., the Thai ancestors were the native AA groups who shifted languages and culture to TK) and continuous migration (i.e., gene flow between a migrant TK and native AA groups) that were developed based on known historical hypotheses (Figure S6). The immigrant and endogenous scenarios postulated an initial split of AA and Dai populations, with a subsequent tree-like split of the target group from Dai (immigrant) or AA (endogenous) populations. The continuous migration model shared the same demographic history as the immigrant model, but also allowed subsequent bidirectional migration between the newly originated population and the AA population. All of the simulations assumed uniform population sizes, fixed separation times based on historical records, a fixed mutation rate of 8.71×10^{-10} substitutions/bp/year (Helgason et al. 2015) and a prior distribution for both effective population sizes and migration rates (Table S3). Finally, due to the uneven sample size between the tested groups, we simulated a number of individuals equal to the lowest sample size among the populations in the model.

We simulated the derived site frequency spectrum (unfolded-SFS) for 2,364,048 loci using the fastsimcoal simulator (Excoffier and Foll 2011) with the flag -s, through the software package ABCtoolbox (Wegmann et al. 2011) and running 50,000 simulations for each model. The observed SFS was calculated with the software 4P (Benazzo et al. 2015). To determine the best performing scenario in each set we employed the model selection procedure ABC-RF (Pudlo et al. 2016), which relies on random forest machine learning methodology (Breiman 2001). This classification algorithm is trained on a reference table of simulations and allows the prediction of the most suitable model at each value of a set of covariates (i.e. the summary statistics). Additional details concerning the ABC-RF analyses are described in our previous study (Kutanan et al. 2018b).

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel, JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. Nature 526(7571): 68-74.

Arias L, Schröder R, Hübner A, Barreto G, Stoneking M, Pakendorf B. 2018. Cultural Innovations Influence Patterns of Genetic Diversity in Northwestern Amazonia. Mol Biol Evol. 35(11): 2719-2735.

Bae CJ, Douka K, Petraglia MD. 2017. Human colonization of Asia in the Late Pleistocene: An introduction to supplement 17. Curr Anthropol. 58: S373–S382.

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. 2012. Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty. Mol. Biol. Evol 29(9): 2157-2167.

Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P. 2013. Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics. Mol. Biol. Evol 30(2): 239-243.

Barbieri C, Ni S, Lippold S, Schröder R, Mpoloka SW, Purps J, Roewer L, Stoneking M, Pakendorf B. 2016. Refining the Y chromosome phylogeny with southern African sequences. Hum Genet. 135: 541–553.

Belwood P. 2018. The search for ancient DNA heads east. Science 361(6397): 31-32.

Benazzo A, Panziera A, Bertorelle G. 2015. 4P: fast computing of population genetics statistics from large DNA polymorphism panels. Ecol Evol. 5(1): 172-175.

Besaggio D, Fuselli S, Srikummool M, Kampuansai J, Castrì L, Tyler-Smith C, Seielstad M, Kangwanpong D, Bertorelle G. 2007. Genetic variation in Northern Thailand Hill Tribes: origins and relationships with social structure and linguistic differences. BMC Evol Biol. 7 Suppl 2 (Suppl 2): S12.

Breiman L. 2001. Random forests. Machine learning 45(1): 5-32.

Browning BL, Browning SR. 2016. Genotype Imputation with Millions of Reference Samples. Am J Hum Genet. 98 (1): 116–26.

Brunelli A, Kampuansai J, Seielstad M, Lomthaisong K, Kangwanpong D, Ghirotto S, Kutanan W. 2017. Y chromosomal evidence on the origin of northern Thai people. PLoS ONE 12(7): e0181935.

Cai X, Qin Z, Wen B, Xu S, Wang Y, Lu Y, Wei L, Wang C, Li S, Huang X, et al. 2011. Human Migration through Bottlenecks from Southeast Asia into East Asia during Last Glacial Maximum Revealed by Y Chromosomes. PLoS ONE 6(8): e24282.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9: 772.

Demeter F, Shackelford LL, Bacon AM, Duringer P, Westaway K, Sayavongkhamdy T, Braga J, Sichanthongtip P, Khamdalavong P, Ponche JL et al. 2012. Anatomically modern human in Southeast Asia (Laos) by 46 ka. Proc Natl Acad Sci USA. 109(36): 14375–14380.

Diffloth G.1984. The Dvaravati Old Mon Language and Nyah Kur. Bangkok: Chulalongkorn University Print House.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. A Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29: 1969–1973.

Excoffier L, Foll M. 2011. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27(9): 1332-1334.

Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 10: 564–567.

Guillou E. 1999. The Mons: A civilization of Southeast Asia. Bangkok: Siam Society Under Royal Patronage.

Hallast P, Batini C, Zadik D, Maisano Delser P, Wetton JH, Arroyo-Pardo E, Cavalleri GL, de Knijff P, Destro Bisol G, Dupuy BM, et al. 2015. The Y-Chromosome Tree Bursts into Leaf: 13,000 High-Confidence SNPs Covering the Majority of Known Clades. Mol Biol Evol. 32(3): 661-673.

Helgason A, Einarsson AW, Guðmundsdóttir VB, Sigurðsson A, Gunnarsdóttir ED, Jagadeesan A, Ebenesersdóttir SS, Kong A, Stefánsson K. 2015. The Y-chromosome point mutation rate in humans. Nat Genet 47: 453–457.

Heller R, Chikhi L, Siegismund HR. 2013. The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History. PLoS ONE 8(5): e62992.

Higham C, Thosarat R. 2012. Early Thailand from prehistory to Sukhothai. Bangkok: River Books.

Higham C. 2013. Hunter-gatherers in Southeast Asia: From prehistory to the present. Hum Biol. 85: 21–43.

Higham C. 2014. Early mainland Southeast Asia: from first humans to Angkor. Bangkok: River Books Press.

Higham C. 2017. First farmers in mainland Southeast Asia. J Indo-Pac Archaeol. 41: 13-21.

Iwai K, Hirono A, Matsuoka H, Kawamoto F, Horie T, Lin K, Tantular IS, Dachlan YP, Notopuro H, Hidayah NI, et al. 2001. Distribution of glucose-6-phosphate dehydrogenase mutations in Southeast Asia. Hum Genet. 108: 445-449.

Jobling M, Hollox E, Kivisild T, Tyler-Smith C. 2004. Agricultural expansions In: Human Evolutionary Genetics. New York. Garland Publishing.

Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11(1): 94.

Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24(11): 1403-1405.

Kumar V, Langstieh BT, Madhavi KV, Naidu VM, Singh HP, Biswas S, Thangaraj K, Singh L, Reddy BM. 2006. Global Patterns in Human Mitochondrial DNA and Y-Chromosome Variation Caused by Spatial Instability of the Local Cultural Processes. PLoS Genet. 2(4): e53.

Karmin M, Saag L, Vicente M, Wilson Sayres MA, Järve M, Talas UG, Rootsi S, Ilumäe AM, Mägi R, Mitt M, et al. 2015. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res. 25: 459–466.

Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 40(1):e3.

Kutanan W, Kampuansai J, Fuselli S, Nakbunlung S, Seielstad M, Bertorelle G, Kangwanpong D. 2011. Genetic structure of the Mon-Khmer speaking groups and their affinity to the neighbouring Tai populations in Northern Thailand. BMC Genet. 12: 56.

Kutanan W, Kampuansai J, Srikummool M, Kangwanpong D, Ghirotto S, Brunelli A, Stoneking M. 2017. Complete mitochondrial genomes of Thai and Lao populations indicate an ancient origin of Austroasiatic groups and demic diffusion in the spread of Tai–Kadai languages. Hum Genet. 136 (1): 85-98.

Kutanan W, Kampuansai J, Changmai P, Flegontov P, Schröder R, Macholdt E, Hübner A, Kangwanpong D, Stoneking M. 2018a. Contrasting maternal and paternal genetic variation of hunter-gatherer groups in Thailand. Sci Reports 8: 1536.

Kutanan W, Kampuansai J, Brunelli A, Ghirotto S, Pittayaporn P, Ruangchai S, Schröder R, Macholdt E, Srikummool M, Kangwanpong D, et al. 2018b. New insights from Thailand into the maternal genetic history of Mainland Southeast Asia. Eur J Hum Genet. 26(6): 898-911.

Lipson M, Cheronet O, Mallick S, Rohland N, Oxenham M, Pietrusewsky M, Pryce TO, Willis A, Matsumura H, Buckley H, et al. 2018 Ancient genomes document multiple waves of migration in Southeast Asian prehistory. Science 361(6397):92-95.

Mallick, S. Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, Skoglund P, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538: 201–206.

Matsuoka H, Nguon C, Kanbe T, Jalloh A, Sato H, Yoshida S, Hirai M, Arai M, Socheat D, Kawamoto F. 2005. Glucose-6-phosphate dehydrogenase (G6PD) mutations in Cambodia: G6PD Viangchan (871G>A) is the most common variant in the Cambodian population. J Hum Genet. 50(9): 468-472.

McColl H, Racimo F, Vinner L, Demeter F, Gakuhari T, Moreno-Mayar JV, van Driem G, Gram Wilken U, Seguin-Orlando A, de la Fuente Castro C, et al. 2018. The prehistoric peopling of Southeast Asia. Science 361(6397): 88-92.

Nahhas RW. 2007. Sociolinguistic survey of Lawa in Thailand. Chiang Mai: Survey Unit Department of Linguistics Faculty of Humanities Payap University.

Nuchprayoon I, Louicharoen C, Charoenwej W. 2008. Glucose-6-phosphate dehydrogenase mutations in Mon and Burmese of southern Myanmar. J Hum Genet. 53(1): 48-54.

Ocharoen S. 1998. Mons in Thailand. Bangkok: Thailand Research Research Fund (In Thai).

Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet. 29(1): 20-21.

Penth H. 2000. A Brief History of Lanna: Civilizations of North Thailand. Chiang Mai: Silkworm Books.

Pon Nya M. 2001. Ethnic identity and political autonomy of the Mon. In McCormick P, Jenny M and Baker C, editors. The Mon over two millennia: Monuments, manuscripts, movements. Bangkok: Institute of Asian Studies, Chulalongkorn University. p. 169- 202.

Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, et al. 2016. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet. 48: 593–599.

Poznik GD. 2016. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men, unpublished data, https://www.biorxiv.org/content/early/2016/11/19/088716, last access 8 May 2018.

Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP. 2016. Reliable ABC model choice via random forests. Bioinformatics 32(6): 859-866.

Renaud G, Stenzel U, Kelso J. 2014. LeeHom: Adaptor trimming and merging for Illumina sequencing reads. Nucleic Acids Res. 42: e141.

Renaud G, Stenzel U, Maricic T, Wiebe V, Kelso J. 2015. deML: Robust demultiplexing of Illumina sequences using a likelihood-based approach. Bioinformatics 31: 770–772.

R Development Core Team. 2016. R: A Language and Environment for Statistical Computing. Vienna R: Foundation for Statistical Computing. http://www.R-project.org/.

Saraya D. 1999. (Sri) Dvaravati: The initial phase of Siam's history. Bangkok: Muang Boran Publishing House.

Shoocondej R. 2006. Late Pleistocene activities at the Tham Lod rockshelter in highland Bang Mapha, Mae Hongson Province, Northwestern Thailand. In: Bacus EA, Glover IC, Pigott VC, editors. Uncovering Southeast Asia's past. Singapore: NUS Press. p. 22–37.

Schliesinger J. 2000. Ethnic groups of Thailand: non-Tai-speaking peoples. Bangkok: White Lotus Press.

Simons GF, Fennig CD. 2018. Ethnologue: Languages of the World. 21th edn. Texas: SIL International.

Sun H, Zhou C, Huang X, Lin K, Shi L, Yu L, Liu S, Chu J, Yang Z. 2013. Autosomal STRs provide genetic evidence for the hypothesis that Tai people originate from Southern China. PLoS ONE 8: e60822.

van Driem GL. 2017. The domestications and the domesticators of Asian rice. In: Robbeets M. Savelyev A. editors. Language Dispersal Beyond Farming. Amsterdam: John Benjamins Publishing Company. p. 183–214.

Wang CC, Huang Y, Yu X, Chen C, Jin L, Li H. 2016. Agriculture driving male expansion in Neolithic Time. Sci China Life Sci. 59: 643-646.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L. 2011. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics 11(1): 116.

Wen S-Q, Tong X-Z, Li H. 2016. Y-chromosome-based genetic pattern in East Asia affected by Neolithic transition. Quat Int. 426: 50-55.

Wilkins JF, Marlowe FW. 2006. Sex-biased migration in humans: what should we expect from genetic data? Bioessays 28(3): 290–300.

Yan S, Wang C-C, Zheng H-X, Wang W, Qin Z-D, Wei L-H, Wang Y, Pan X-D, Fu W-Q, He Y-G, et al. 2014. Y Chromosomes of 40% Chinese Descend from Three Neolithic Super-Grandfathers. PLoS ONE 9(8): e105691.

Acknowledgements

We would like to thank all sample donors, village chief and coordinators, i.e. Sukhum Ruangchai, Khamnikone Sipaseuth, Worasitikulya Taratima, Saksuriya Triyarach, Narongdech Mahasirikul, Supada Khonyoung, Dusit Boonmekam, Tharanat Hin-on, Kantaphon Chueahor, Pittayawat Pittayaporn and Waraporn Hongsaphinan for assistance in collecting samples. We thank Murray Cox, Brigitte Pakendorf and Rasmi Shoocondej for valuable discussion. This study was supported by the Max Planck Institute for Evolutionary Anthropology. WK was also funded by the Thailand Research Fund (Grant number RSA6180058), Khon Kaen University (Grant number 6100100) and KKU's Thai Visiting Scholar 2018. MSr was funded by Naresuan University (Grant number R2561B029).

Author contributions

W.K. and M.S. conceived and designed the project; W.K., J.K. and M.Sr. collected samples; W.K. and R.S. generated data; W.K., A.B., S.G., L.A., A.H. and E.M. involved data analyses; W.K. and M.S. wrote the paper with input from all co-authors.

Table 1 AMOVA results. The numbers in parentheses show the percent variation of MSY by excluding the Maniq (MN) and asterisks indicate significant level ($P < 0.01$).

Groups	Number of groups	Number of populations	Percent variation					
			Within populations		Within groups		Among groups	
			MSY	mtDNA	MSY	mtDNA	MSY	mtDNA
Total	1	59 (58)	88.88 (89.46)	91.51	11.12* (10.54*)	8.55*		
Language	3	59 (58)	88.21* (98.05*)	91.20*	10.16* (1.96*)	8.18*	1.63* (-0.01)	0.62*
Austroasiatic	1	24 (23)	79.99 (81.51)	85.97	20.01* (18.49*)	14.03*		
Mon	1	7	96.08	93.10	3.92*	6.90*		
Htin	1	3	88.47	74.29	11.53*	25.71*		
Lawa	1	3	65.57	92.22	34.43*	7.78*		
Sino-Tibetan (Karen)	1	4	97.71	93.49	2.29	6.51*		
Tai-Kadai	1	31	95.52	95.67	4.48*	4.33*		
Central Thai	1	7	98.53	98.36	1.47	1.64*		
Khon Mueang	1	4	101.83	95.80	-1.83	4.20*		
Lao Isan	1	4	98.16	97.69	1.84	2.31*		
Geography	6 (5)	59 (58)	88.27* (98.07*)	91.40*	9.35* (2.02*)	8.40*	2.38* (-0.09)	0.20*
Northern	1	26	85.51	88.84	14.49*	11.16*		
Northeastern	1	16	96	91.29	8.00*	8.71*		
Central	1	11	94.61	95.86	5.39*	4.14*		
Western	1	3	93.97	99.11	6.03*	0.89		

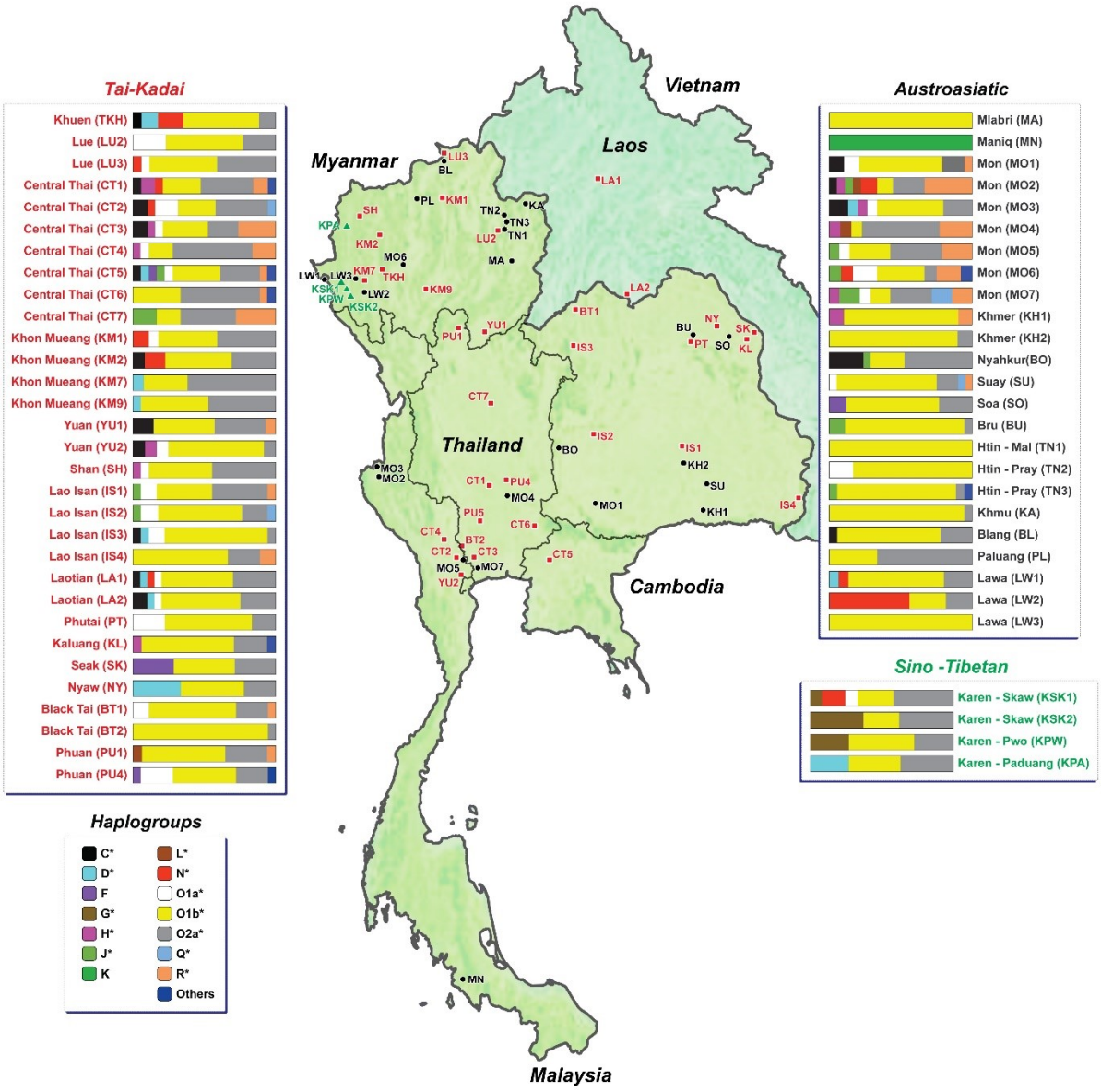


Figure 1 Map showing sample locations and haplogroup distributions.

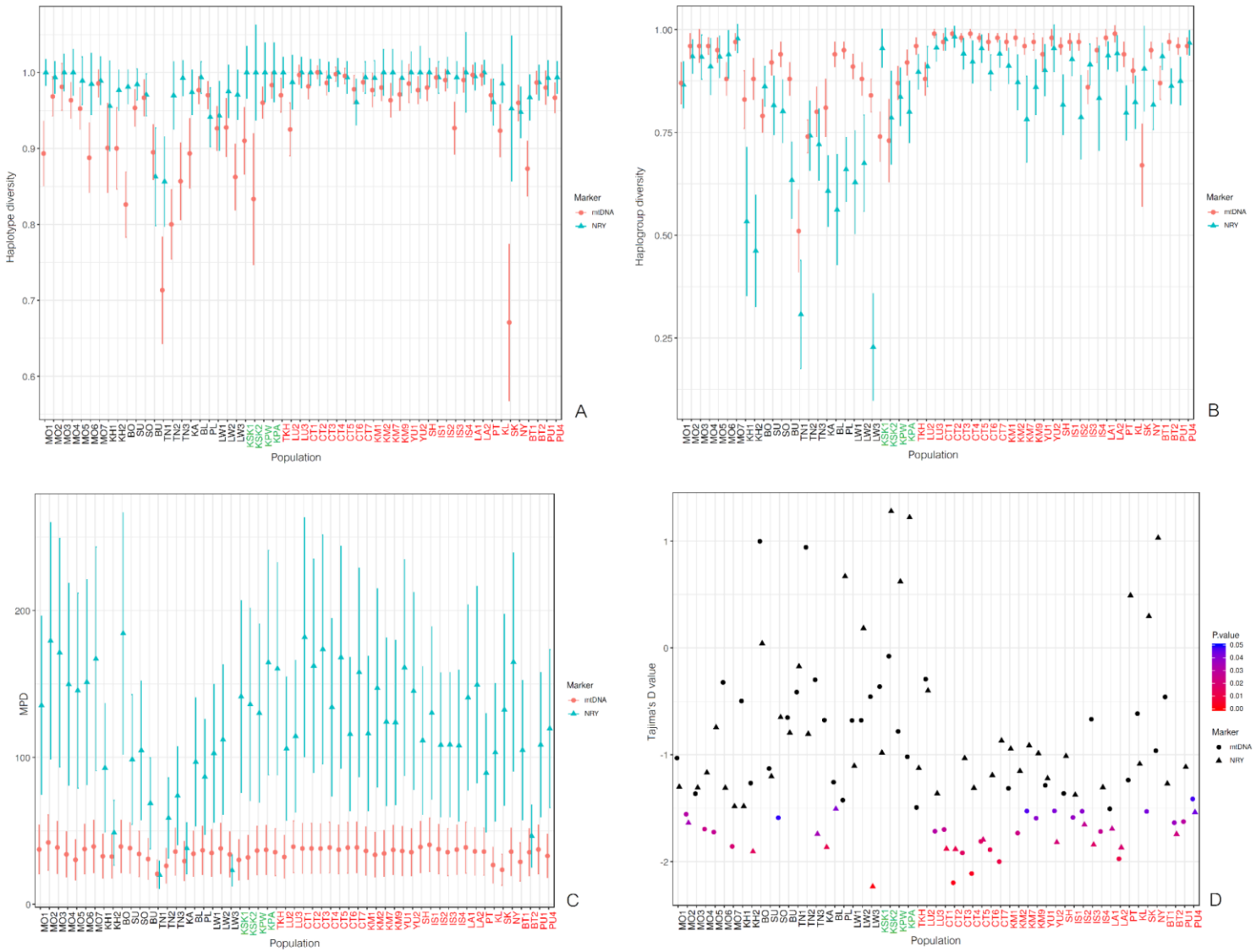


Figure 2 Genetic diversity values of MSY and mtDNA in the studied populations, excluding the Maniq (MN) and Mlabri (MA): haplotype diversity (A), haplogroup diversity (B), mean number of pairwise difference (MPD) (C), and Tajima D's values (D). More information and all genetic diversity values are provided in Table S5.

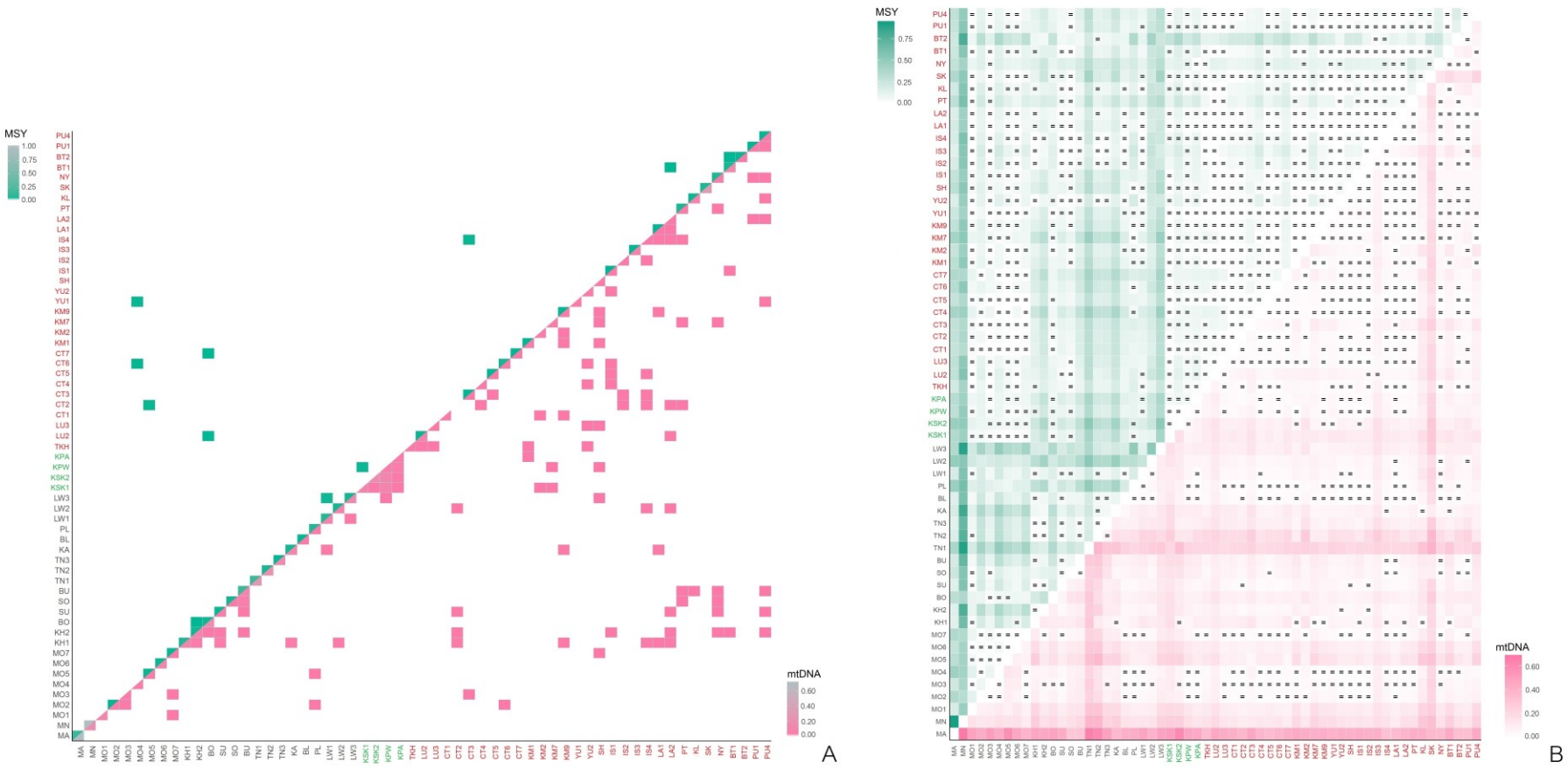


Figure 3 Relative shared haplotypes (A) and heat plot of Φ_{st} (B) between studied populations for the MSY and for mtDNA.

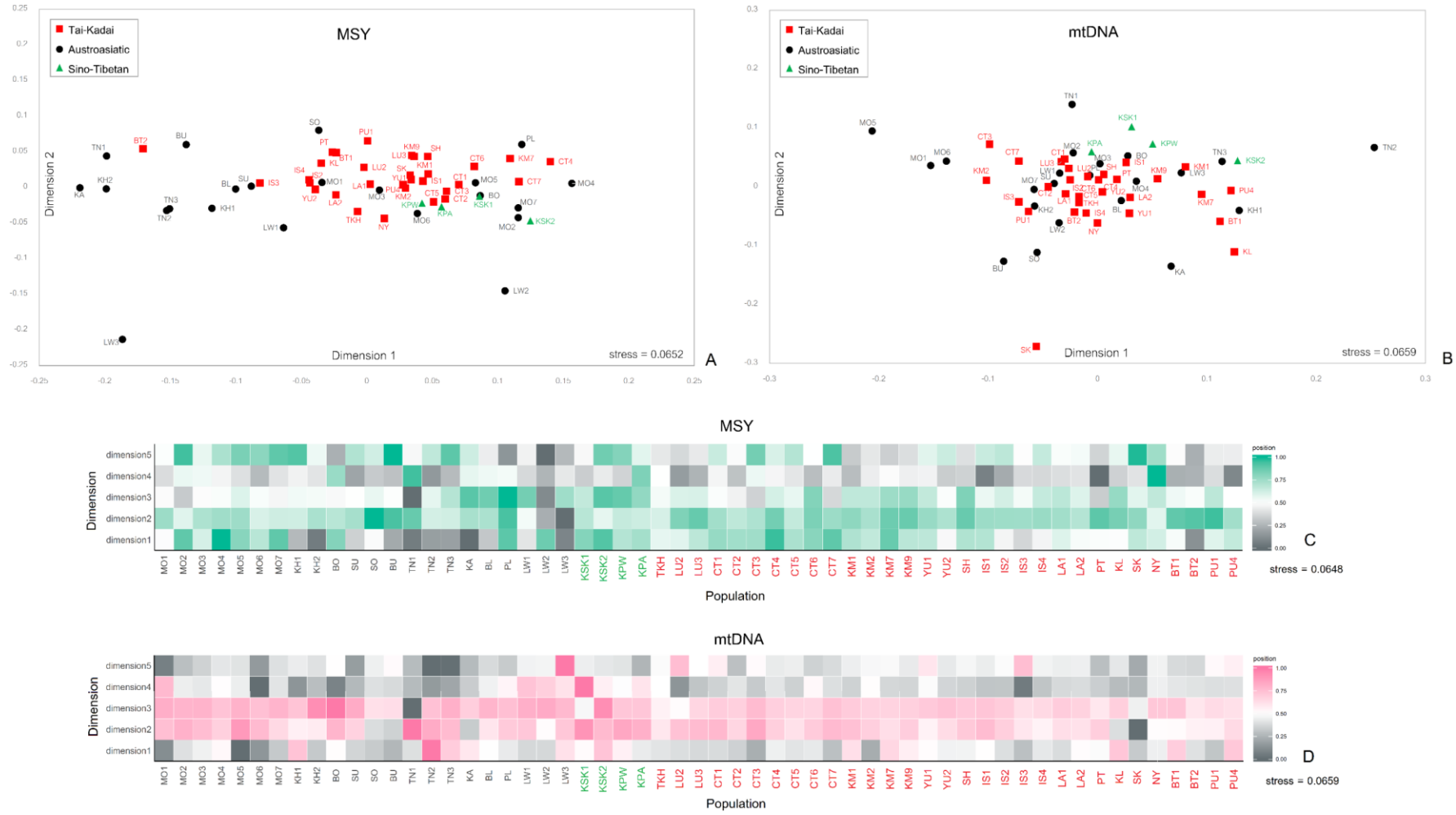


Figure 4 The two-dimensional MDS plot and five-dimensional MDS heat plot based on the Φ_{st} distance matrix for 57 populations (after removal of Maniq and Mlabri) of MSY (A and C) and mtDNA (B and D).

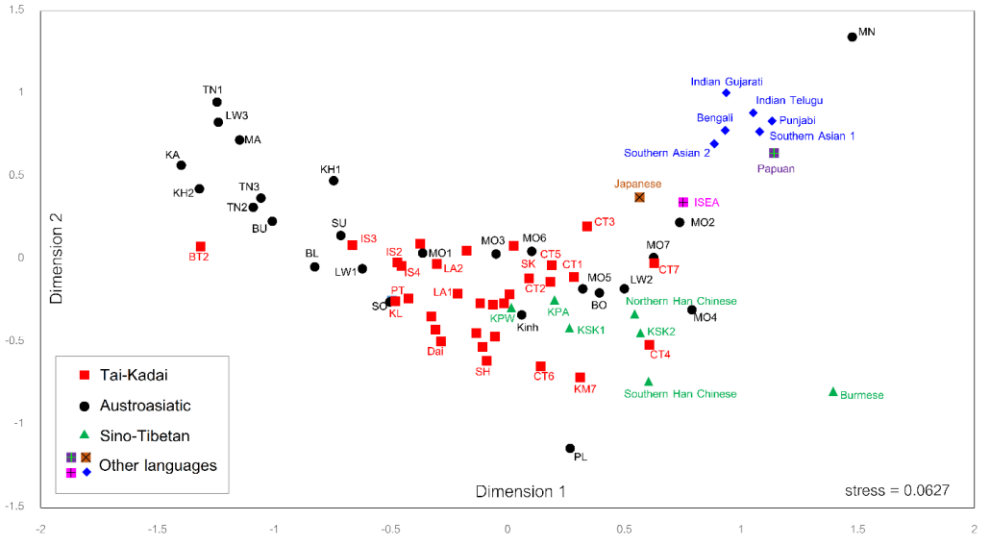


Figure 5 The two-dimensional MDS plot based on the MSY Φ_{st} distance matrix for 73 populations. Population details are listed in Figure 1 and Tables S5 and S7.

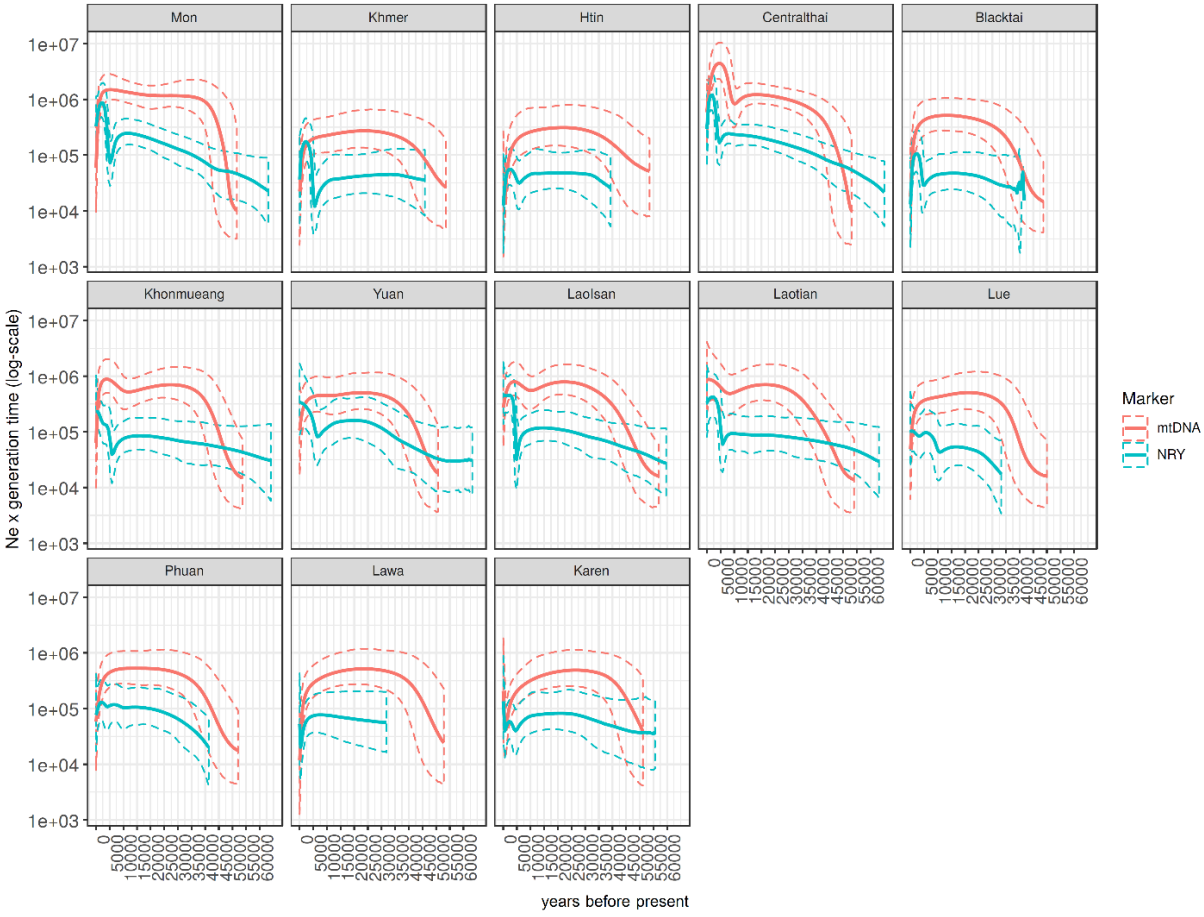


Figure 6 The BSP plots based on the MSY and mtDNA of 13 ethnicities from Thailand and Laos; Mon, Khmer, Htin, Central Thai, Black Tai, Khon Mueang, Yuan, Lao Isan, Laotian, Lue, Phuan, Lawa, Karen. Solid lines are the median estimated effective population size (y-axis) through time from the present in years (x-axis). The 95% highest posterior density limits are indicated by dotted lines.

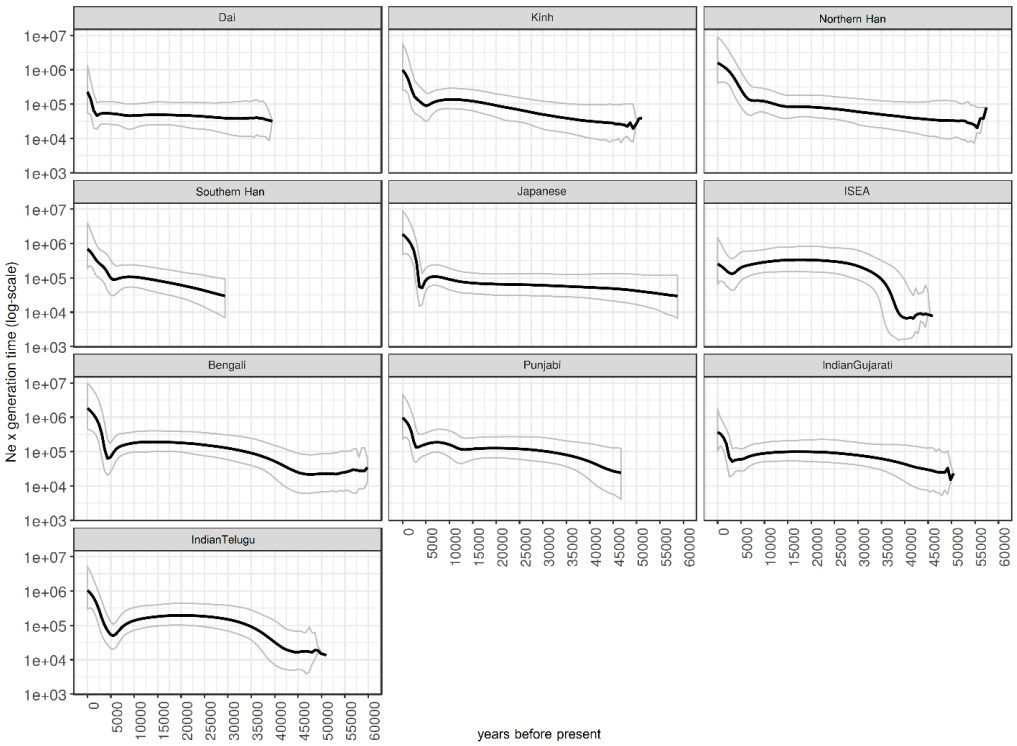


Figure 7 The BSP plots of Asian populations. Solid lines are the median estimated paternal effective population size (y-axis) through time from the present in years (x-axis). The 95% highest posterior density limits are indicated by dotted lines.

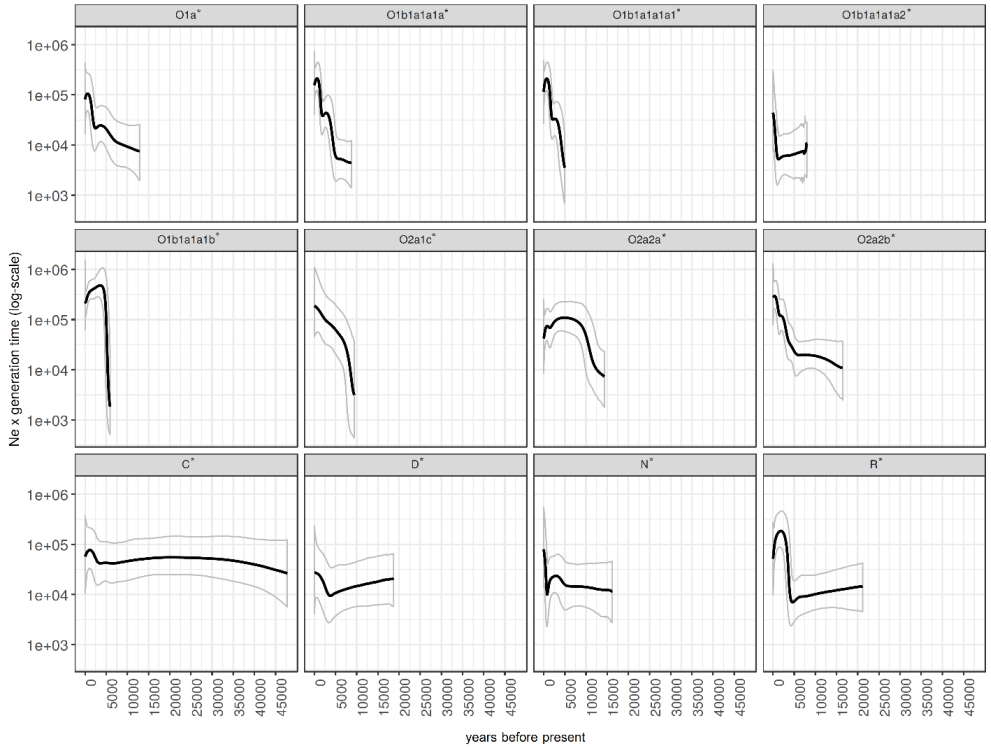


Figure 8 The BSP plots for each major haplogroup. Solid lines are the median estimated paternal effective population size (y-axis) through time from the present in years (x-axis). The 95% highest posterior density limits are indicated by dotted lines.

Supplementary Text

Genetic relatedness among populations

The MA and MN show large differences from the other populations in the heat plots of Φ_{st} values (Figure S5). However, in general both MSY and mtDNA results show relatively larger genetic heterogeneity of the AA groups vs. genetic homogeneity of the TK and ST groups (Figure S3 and S5). After excluding these MA and MN as outliers, the first dimension of the plot divides the AA populations into two groups: one is diverged from the TK cloud, i.e. KH, KA, SU, TN, BU, BL, LW1 and LW3 and another is interspersed with the TK groups, i.e. SO, MO, BO, and PL (Figure 4A). The MDS heat plot for the MSY supported the divergence of AA populations and also emphasized the similarity between some AA populations and TK populations (Figure 4C). The ST speaking-Karen populations are close to the AA-speaking Mon in the right side of the plot (Figure 4A). Among the TK-speaking populations, BT2 and IS3 are closer to the AA groups on the left side of the plot while the central Thai (CT1-CT7) and one Khon Mueang group (KM7) are closer to the AA groups (MO, PL and LW2) and Karen on the right side (Figure 4A), in agreement with the MDS heat plot for the MSY (Figure 4C). In the second dimension, the Lawa groups are very differentiated (Figure 4A), in accordance with the AMOVA (Table 1) and heat plot results Figure 4C). The heat plot of MSY Φ_{st} values supports strong genetic homogeneity in the TK and ST groups and also generally shows non-significant differences between the Mon (all groups) and the TK populations, especially with the CT groups, which are different from the other AA speaking populations (Figure 4C). For the MDS of mtDNA (Figure 4B), the Mon generally showed genetic affinity with the TK groups in the center of the plot, with the exception of MO1, MO5 and MO6, which differ from the other Mon groups, as can be also seen in the MDS plot (Figure 4B) and heat plot (Figure 4D). Moreover, contrasting relationships based on the MSY vs. mtDNA was observed for the SK and SO from northeastern Thailand, and BT2 from central Thailand. The mtDNA differentiation from other populations was stronger than that for the MSY for the SK and SO, while opposite was observed for the BT2 (Figure 4A and 4B).

Thai MSY haplogroup distribution

Among the 928 MSY sequences from Thailand, there are 92 specific haplogroups. Because some of these are subhaplogroups of other haplogroups, we use the following nomenclature: an asterisk denotes a parent haplogroup and all subhaplogroups, while the lack of an asterisk denotes just that specific haplogroup. O1b* is by far the most frequent haplogroup (51.19%) and is present in all populations except the MN, who have only haplogroup K (Figure 1; Table S1). There are two subclades of O1b*: O1b1a1* or O-PK4*

(99.37%), and O1b1a2 or O-Page59 (0.63%). Only O1b1a1* was previously reported to be wide spread in northern Thailand (Brunelli et al., 2017), while O1b1a2 is newly reported here, occurring in CT5, YU2 and PU4 (Table S1). There are several subclades of O1b1a1*; the most frequent (50.54%) is O1b1a1a* or O-M95*, which occurs in almost all populations. However, almost half of the AA groups show a very high frequency of O-M95* (>70%), i.e. KH1-KH2, KA, BU, BL, SU, TN1-TN3, MA and LW3 (Figure 1; Table S1) while only two TK populations, i.e. BT2 (94.44%) and IS3 (72.22%) have a high frequency of O-M95*. It appears that the frequency of O-M95* is a major driver of the patterns in the MDS plot in dimension 1 (Figure 4A): we find that O-M95* is at high frequency in the populations on the left of the plot and gradually decreases to very low frequency in the populations on the right side, e.g. MO2, MO4, MO7, BO, CT4 and CT7 (Figure 4A).

O-M95* has also been reported to be frequent in AA populations from Cambodia (Zhang et al., 2015) and Laos (Cai et al., 2011), but infrequent in populations from southern China (Zhang et al., 2015) and rare elsewhere in MSEA (Trejaut et al., 2014). The CA analysis (based on haplogroup frequencies) also supports the divergence of these AA populations, with many O1b* sublineages, e.g. O1b1a1a1b1a (O-B426) and O1b1a1a1a1a* (O-F2758*) (Figure S1). With a total frequency at 11.10%, O1b1a1a1b1a* is prevalent in the LW3 (88.23%), BL (66.67%), LW1 (60.00%), KA (55.56%), and TN3 (47.06%) populations (Table S1).

O2a* or O-M324* is the second most frequent haplogroup with an overall frequency of 25.86%; this haplogroup has relatively high frequency (>50%) in several populations: MO4 (53.85%), PL (66.67%), CT4 (55.55%), CT6 (55.55%), and KM7 (61.53%); and moderate frequencies in some Mon groups (MO5: 35.71%), some TK groups (KM9 (47.05%), LU3 (41.17%), and SH (44.44%), and all ST speaking Karen (KSK1 (41.67%), KSK2 (37.5%), KPA (36.36%)) (Figure 1, Table S1). Interestingly, we also observe a cline in O2a* frequency in the first dimension of the MDS plot that runs opposite to the O-M95* cline: O2a* is at higher frequency in populations located on the right of the plot and decreases in frequency toward the left side (Figure 4A).

Within O2a*, lineage O2a2b1a1a* or O-F8*, which is equivalent to O-M133*, is the most frequent (13.79% total frequency) and occurs in almost all populations, with fairly high frequencies in PL (50%), KM7 (46.15%), CT4 (38.88%), SH (38.88%), KSK2 (37.50%) and KPA (36.36%) (Table S1). O-M133 has been reported in Han Chinese from Taiwan and Thai from Bangkok (12.73-21.59%) (Trejaut et al., 2014), Northern Han (11.36%), Southern Han (9.61%), Kinh (8.70%), Japanese (9.09%) and Dai (24.44%) (Poznik et al., 2016), but is very rare in Malaysia, Indonesia, the Philippines and South Asia (Trejaut et al., 2014; Poznik et al., 2016).

The last subhaplogroup of O observed in our study is O1a* or O-M119*. With a total frequency of 4.53%, O1a* is prevalent in three TK-speaking populations, i.e. LU2 (23.08%), PU4 (22.22%) and PT (22.22%) and occurs at low frequency in several populations, including central Thai (CT2-CT5), Laotian (LA1-LA2) and Lao Isan (IS1-IS3) (Figure 1; Table S1). O-M119* is thus spread across many TK-speaking groups, and also occurs at high frequency in Austronesian populations (Trejaut et al., 2014), indicating shared genetic lineages between TK and AN speaking groups. O-M119* occurs sporadically in just a few AA groups (Mon, SU and TN2) and at low frequency, in agreement with a previous study of Laos (Cai et al., 2011). The observed O-M119* sequences in the AA groups thus might reflect contact with TK groups.

Overall, the SEA-specific O1b* and O2a* haplogroups (with several sublineages) differentiate our studied populations into at least two main paternal sources, and the frequencies of these two haplogroups correspond to the major differentiation in the MDS plot (Figure 4A). However, there are also several minor non-SEA MSY lineages which promote divergence for some populations, e.g. the Lawa groups. Haplogroup N*, a sister clade of O*, is reported to be distributed in north Asian, Tibeto-Burman, and AA groups in southwestern China (Shi et al., 2013). Only one sublineage (N1c2b2 or N-L665) was found in this study (total frequency of 2.80%) and one third of N-L665 was restricted to LW2, enhancing the divergence of this population. It also sporadically occurs in some AA groups (MO2, MO6 and LW1), Karen (KSK1) and TK groups (THK, LU3, CT1, CT2, KM1, KM2 and LA1) (Table S1).

We also observed some minor haplogroups that are abundant in South/Central Asia (Lippold et al., 2014; Karmin et al., 2015; Poznik et al., 2016), e.g. R*, H*, and J*, occurring at total frequencies of 4.18%, 1.40% and 1.62%, respectively (Figure 1). Haplogroup R* is observed in all Mon groups (41.46% of R) except for MO3, and is at highest frequency in MO2 (33.33%) (Table S1). The same proportion of this haplogroup (41.46% of R*) is also detected in all central Thai groups, except for CT2, and is at high frequency in CT3 (26.32%) and CT7 (27.78%), providing more support for genetic connections between Mon and Central Thais. The remaining proportion (17.02%) of R* was found sporadically (only single samples each) in KH1, SU, YU1, IS1, IS4, BT1 and PU1, probably reflecting recent admixture/gene flow. In agreement with these observations, the CA plots show a correspondence between R1a1a1b (R-Z647) and some central Thai and Mon groups (Figure S1). Haplogroup H* shows a similar haplogroup distribution, i.e. occurring in both some Mon (MO2-MO4 and MO7) and some central Thai (CT1 and CT3-CT5) groups. Haplogroup H* also occurs sporadically in the KH1, YU2, SH and KL groups. Elsewhere, haplogroup H* is found in Burmese and Malayan populations (Karmin et al., 2015) and Central Asian populations (Lippold et al., 2014). Haplogroup J2* is distributed in AA speaking Mon (MO2, MO5, MO6 and MO7), BO, BU, TN3 and TK speaking central Thai (CT5 and CT7) and Lao Isan (IS1 and IS2). Haplogroup J2* has been found in many populations from Central Asia (Lippold et al., 2014). Generally, the haplogroup profile indicates genetic affinities between the

Mon and South/Central Asian groups, which is consistent with the MDS plots (Figure 4A) and results from mtDNA analyses (Kutanan et al., 2017; Kutanan et al., 2018b).

The other minor haplogroups observed in this study are D1*, G1b and F. Haplogroup D1* was found with the highest frequency in NY (33.33%) followed by KPA (27.27%), and occurs at lower frequency in a few other groups (Table S1). Subclade D1a1a (D-N1) is prevalent in Tibetan groups and ST-speaking Qiang in southwestern China, and less prevalent in MSEA (Shi et al., 2008; Wang et al., 2013). Haplogroup G1b (G-L835/L830) was restricted to the Karen, where it was found in three of the four Karen groups. G1* was previously reported to occur in Southwest and Central Asia (Balanovsky et al., 2015). The CA analysis also supports the divergence of the Karen (KSK1, KSK2 and KPW) based on G1b, the divergence of SK based on F, and the differentiation of NY and KPA based on D1a1a (Figure S1). In general, the occurrence of both SEA specific haplogroups and haplogroups prevalent in North Asia/Tibet and Southwest Asia in the Karen suggest multiple parental sources, in agreement with previous studies based on mtDNA (Kutanan et al., 2014; Kutanan et al., 2018b). Haplogroup F (F-M89) was distributed at low frequency in SO, SK and PU4, who migrated from Vietnam during historical times (Schliesinger, 2000) and was also reported in one Kinh sample from Vietnam (Poznik et al., 2016) and five Lahu samples from southern China (Lippold et al., 2014). The origins of this haplogroup are uncertain but it might have originated in the area of present-day Vietnam and southern China; additional studies of Vietnamese populations would be informative.

Reference

Balanovsky O, Zhabagin M, Agdzhoyan A, Chukhryaeva M, Zaporozhchenko V, Utevska O, Highnam G, Sabitov Z, Greenspan E, Dibirova K, et al. 2015. Deep Phylogenetic Analysis of Haplogroup G1 Provides Estimates of SNP and STR Mutation Rates on the Human Y-Chromosome and Reveals Migrations of Iranic Speakers. PLoS ONE 10(4): e0122968.

Brunelli A, Kampuansai J, Seielstad M, Lomthaisong K, Kangwanpong D, Ghirotto S, Kutanan W. 2017. Y chromosomal evidence on the origin of northern Thai people. PLoS ONE 12(7): e0181935.

Cai X, Qin Z, Wen B, Xu S, Wang Y, Lu Y, Wei L, Wang C, Li S, Huang X, et al. 2011. Human Migration through Bottlenecks from Southeast Asia into East Asia during Last Glacial Maximum Revealed by Y Chromosomes. PLoS ONE 6(8): e24282.

Karmin M, Saag L, Vicente M, Wilson Sayres MA, Järve M, Talas UG, Rootsi S, Ilumäe AM, Mägi R, Mitt M, et al. 2015. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res. 25: 459–466.

Kutanan W, Srikummool M, Pittayaporn P, Seielstad M, Kangwanpong D, Kumar V, Prombanchachai T, Chantawannakul P. 2015. Admixed origin of the Kayah (Red Karen) in Northern Thailand Revealed by Biparental and Paternal Markers. Ann Hum Genet. 7: 108–122.

Kutanan W, Kampuansai J, Srikummool M, Kangwanpong D, Ghirotto S, Brunelli A, Stoneking M. 2017. Complete mitochondrial genomes of Thai and Lao populations indicate an ancient origin of Austroasiatic groups and demic diffusion in the spread of Tai–Kadai languages. Hum Genet. 136 (1): 85-98.

Kutanan W, Kampuansai J, Brunelli A, Ghirotto S, Pittayaporn P, Ruangchai S, Schröder R, Macholdt E, Srikummool M, Kangwanpong D, et al. 2018b. New insights from Thailand into the maternal genetic history of Mainland Southeast Asia. Eur J Hum Genet. 26(6): 898-911.

Lippold S, Xu H, Ko A, Li M, Renaud G, Butthof A, Schröder R, Stoneking M. 2014. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. Investigative Genet. 5: 13.

Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, et al. 2016. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet. 48: 593–599.

Schliesinger J. 2000. Ethnic groups of Thailand: non-Tai-speaking peoples. Bangkok: White Lotus Press.

Shi H, Zhong H, Peng Y, Dong YL, Qi XB, Zhang F, Liu LF, Tan SJ, Ma RZ, Xiao CJ, et al. 2008. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. BMC Biol. 6: 45.

Shi H, Qi X, Zhong H, Peng Y, Zhang X, Ma RZ, Su B. 2013. Genetic Evidence of an East Asian Origin and Paleolithic Northward Migration of Y-chromosome Haplogroup N. PLoS ONE 8(6): e66102.

Trejaut JA, Poloni ES, Yen J-C, Lai YH, Loo JH, Lee CL, He CL, Lin M, et al. 2014. Taiwan Y-chromosomal DNA variation and its relationship with Island Southeast Asia. BMC Genet. 15: 77.

Wang C-C, Li H. 2013. Inferring human history in East Asia from Y chromosomes. Investigative Genet. 4: 11.

Supplementary Figures

Figure S1 Correspondence Analysis (CA) results based on haplogroup frequency of 58 populations excluding Maniq (MN). Population abbreviations are shown in Figure 1 and Table S5.

Figure S2 Percent variation among populations in various linguistic or geographic categories, calculated by AMOVA.

Figure S3 The DAPC results based on ethnicity, population, geography and language (A, C, E and G, respectively). The DAPC results, excluding the Maniq based on ethnicity, population, geography and language (B, D, F and H).

Figure S4 The bar plot graphs of within population genetic variation values, i.e. haplotype diversity (A), MPD (B) and haplogroup diversity (C) in patrilocal and matrilocal groups. The shaded bar in each group indicates the mean diversity in each group.

Figure S5 The MDS plot and associated heat plot based on the Φ_{st} distance matrix calculated from the dataset for 59 populations, for the MSY (A) and mtDNA (B). Population abbreviations are in Figure 1and Table S5.

Figure S6 Three demographic models for the ABC analysis (demic diffusion, cultural diffusion and continuous migration). A, B and C represent the different populations and Test 1-5 are the different datasets used in each test. T1, T2 and T3 are either divergence time or time of gene flow.

Figure S7 PCA (Principle Component Analysis) analysis based on Dimension 1 and 2 and Dimension 3 and 4 for the fit between the observed data and the simulated data generated by each model for the origin of Northern Thai (Test 1), Laotian and Lao Isan (Test 2), Laotian (Test 3), Lao Isan (Test 4) and Central Thais (Test 5).

Figure S8 Graphs representing the posterior distribution of each estimated effective population sizes over the extent of the prior range (solid black) and the prior distribution of each parameter (light gray) in each model for the origin of Northern Thai (Test 1), Laotian and Lao Isan (Test 2), Laotian (Test 3), Lao Isan (Test 4) and Central Thais (Test 5).

Supplementary Tables (excel file)

Table S1 Haplogroup frequency.

Table S2 Votes assigned to each model by the Random Forest procedure and posterior probability for the selected model in the ABC analysis.

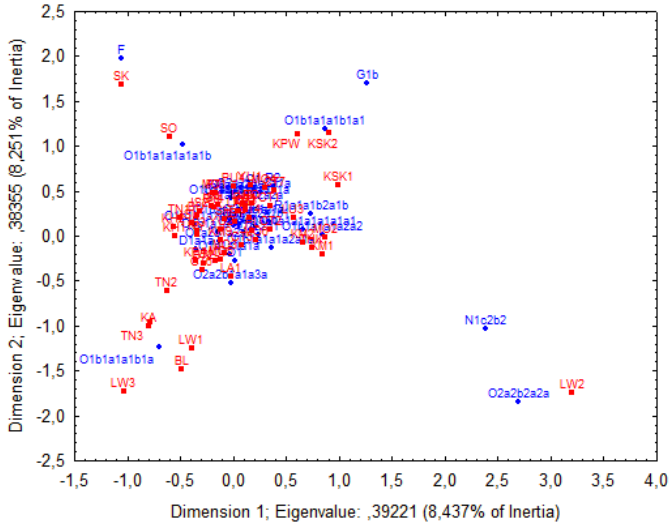
Table S3 Parameters estimation for the selected model in each ABC analysis tested.

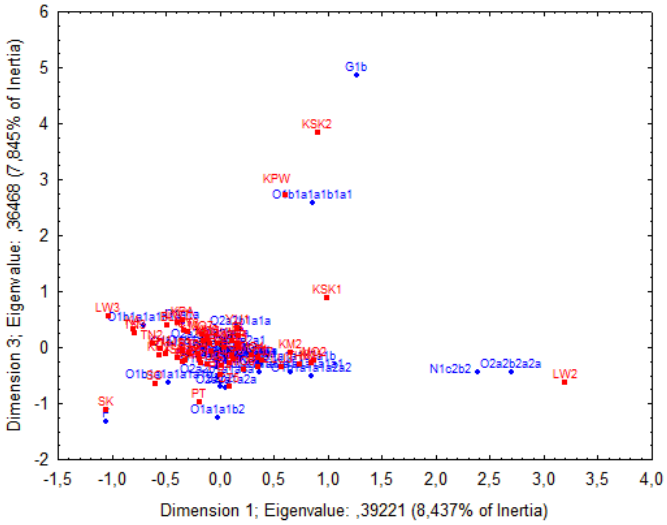
Table S4 Genetic differences (Φst and corrected pairwise differences) between each groups of population used in ABC testes. Numbers in parentheses indicate P-values.

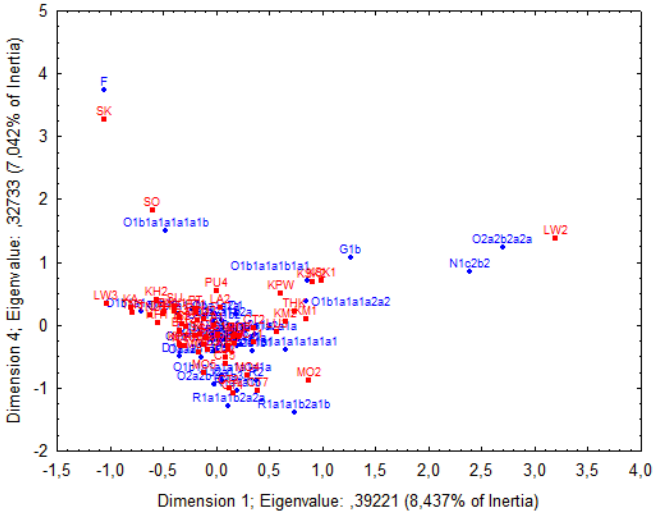
Table S5 General information and genetic diversity values of the studied populations.

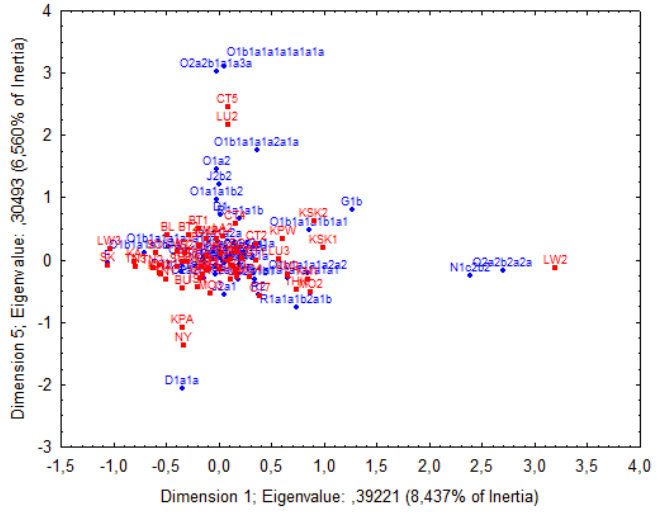
Table S6 MSY probe set details.

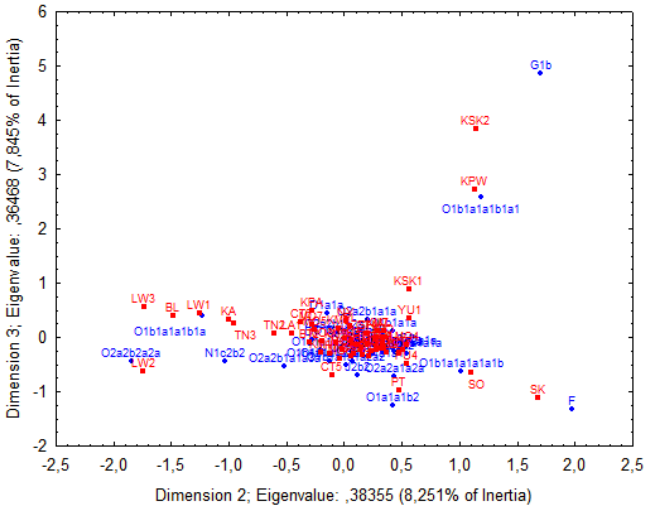
Table S7 Details for the compared populations for MSY data.

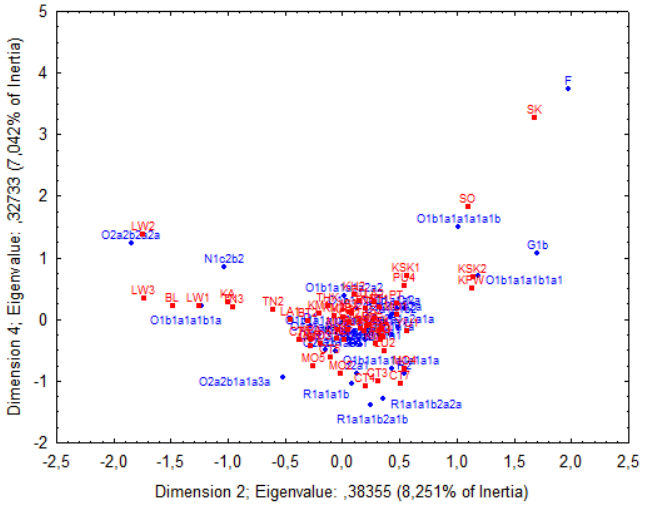












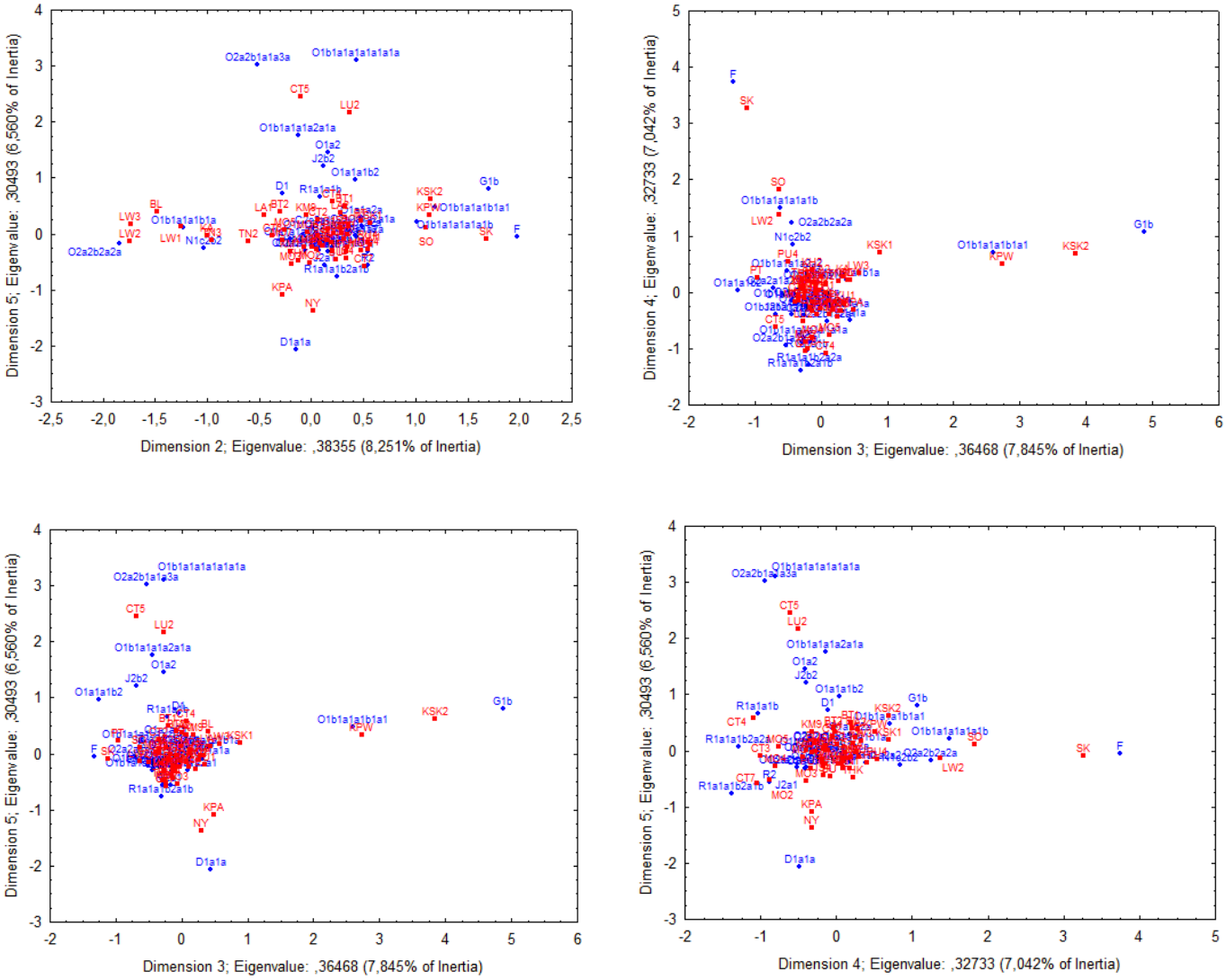


Figure S1 Correspondence Analysis (CA) results based on haplogroup frequency of 58 populations excluding Maniq (MN). Population abbreviations are shown in Figure 1 and Table S5.

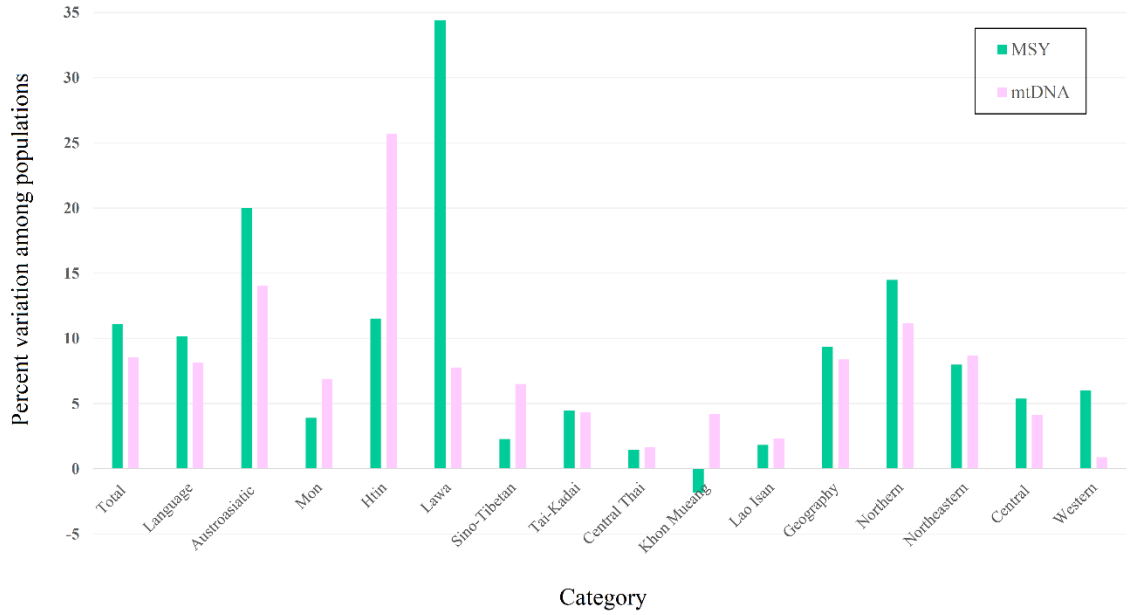
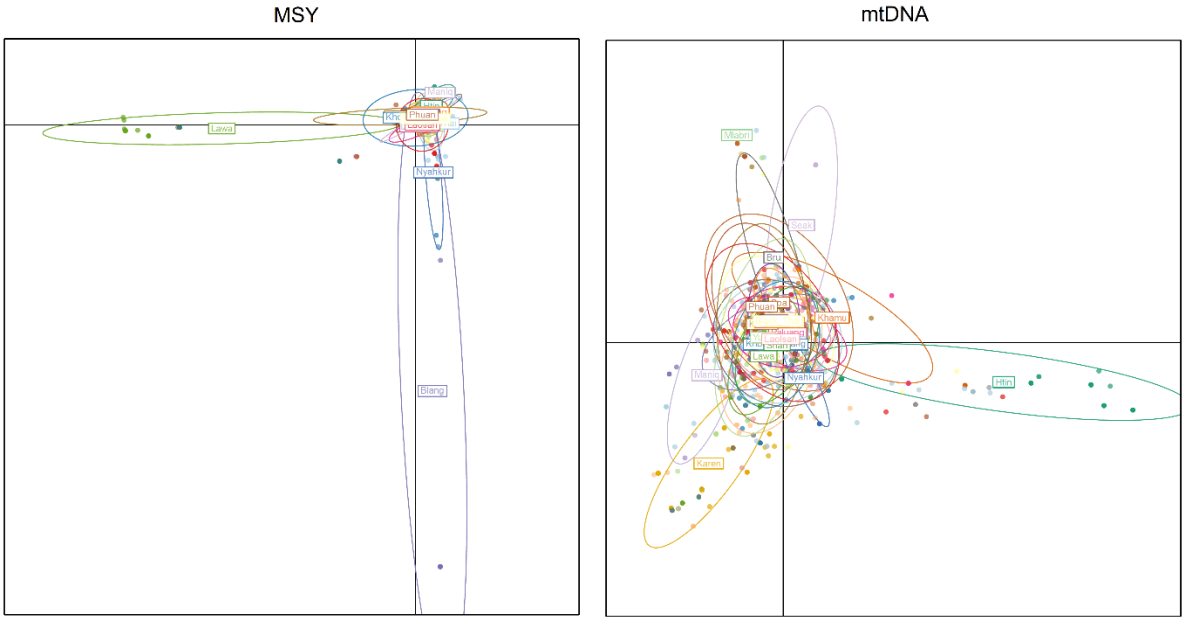
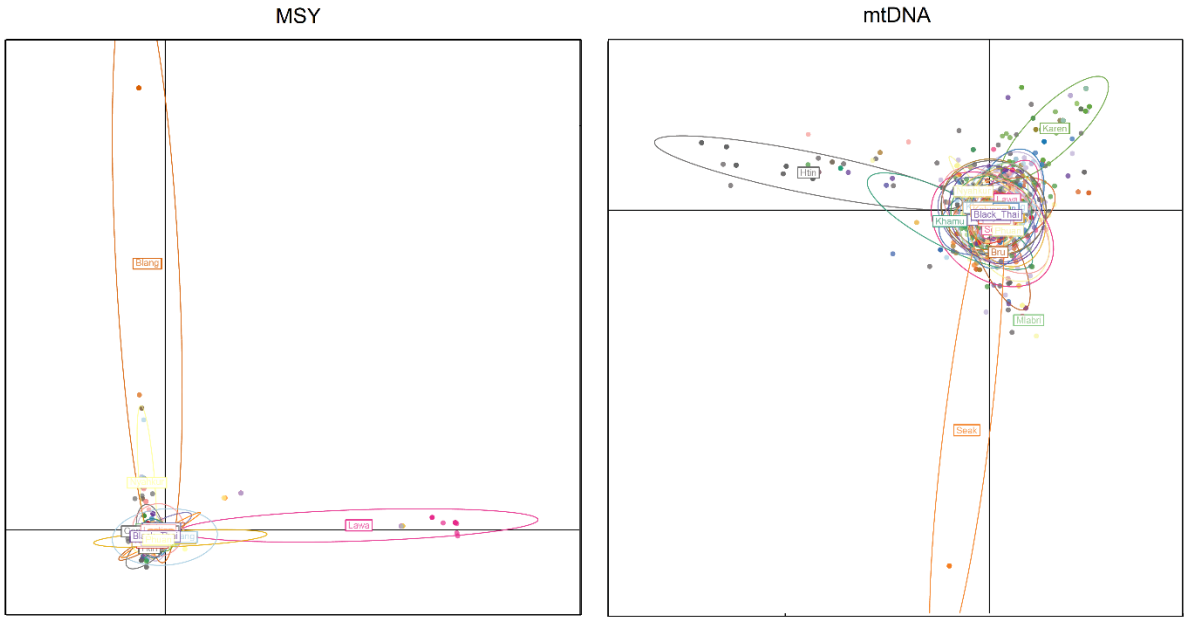


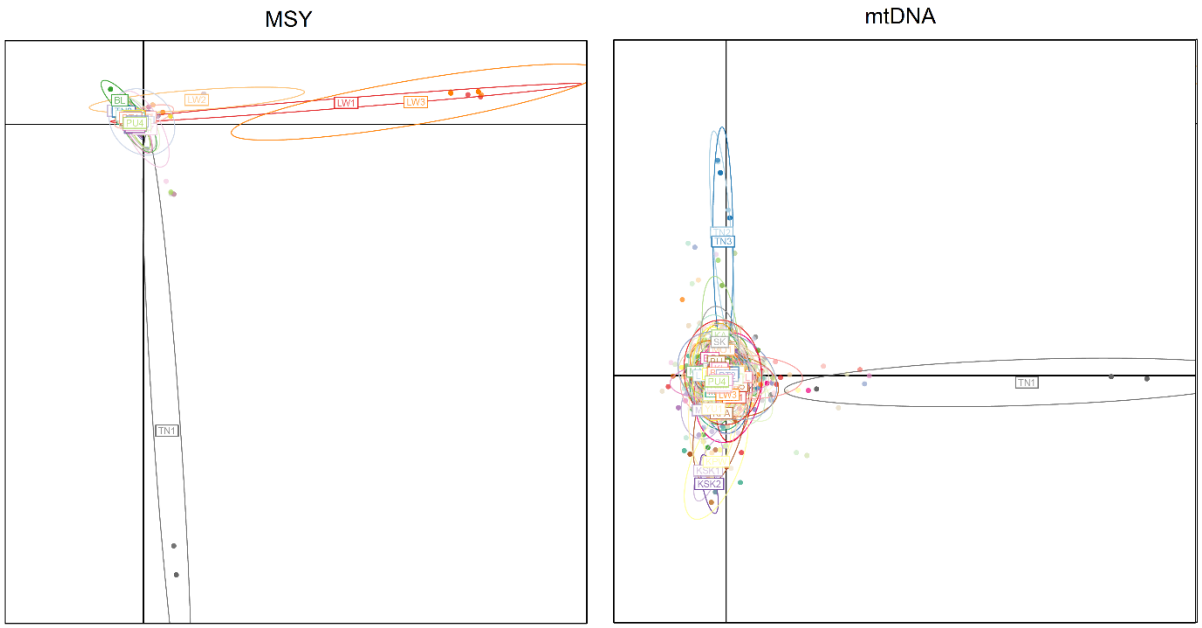
Figure S2 Percent variation among populations in various linguistic or geographic categories, calculated by AMOVA.



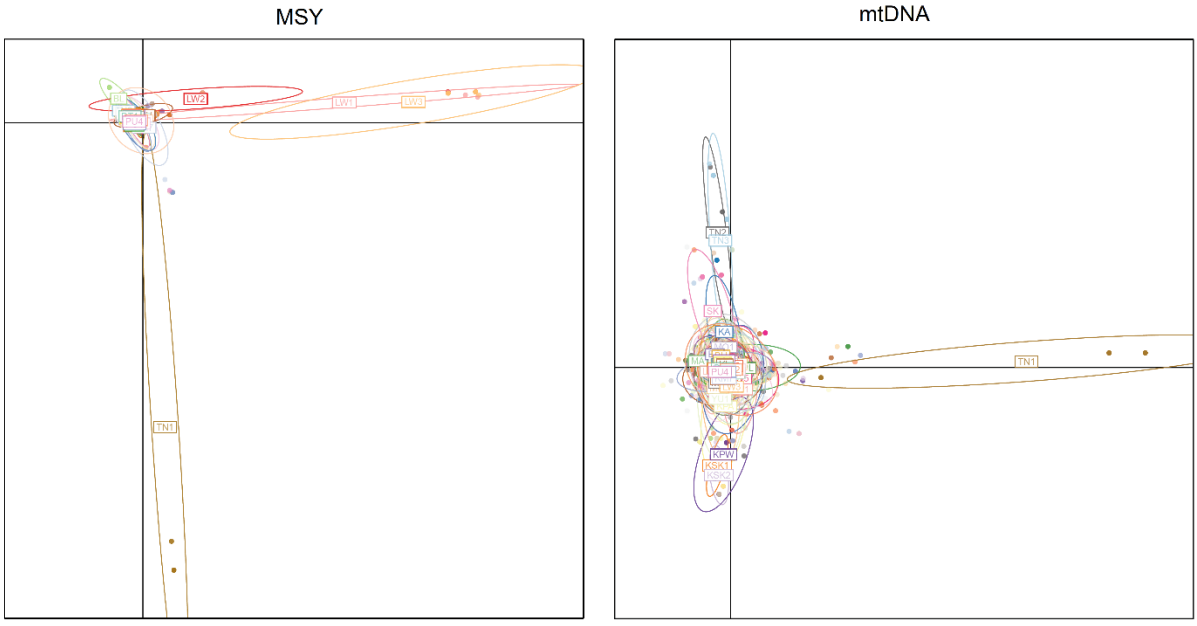
(A)



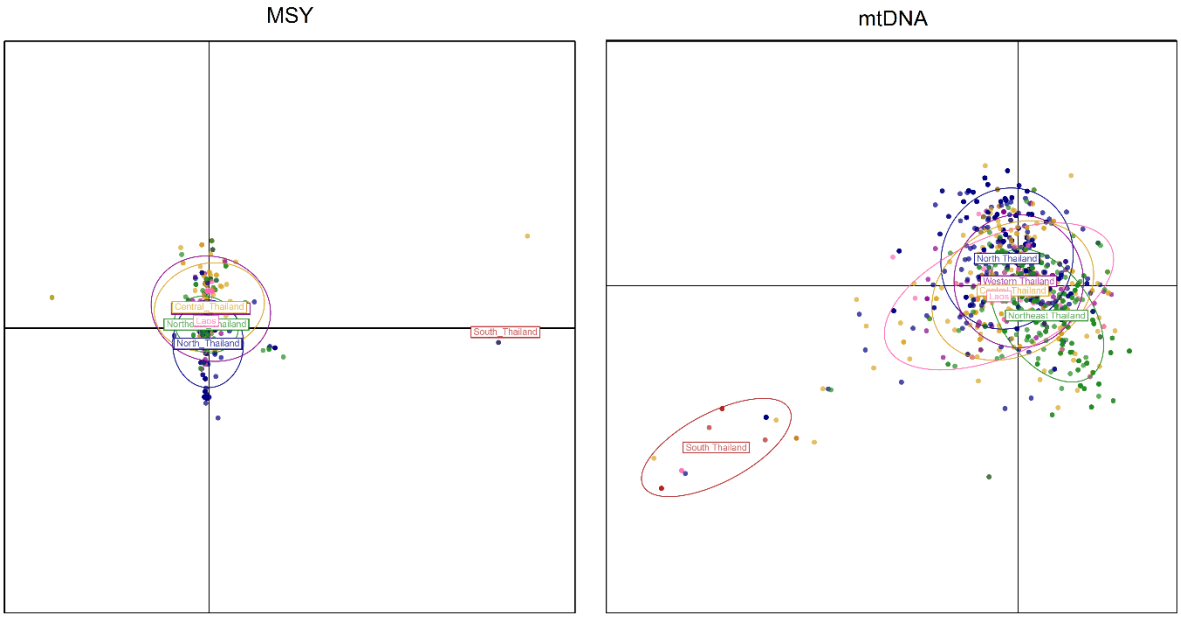
(B)



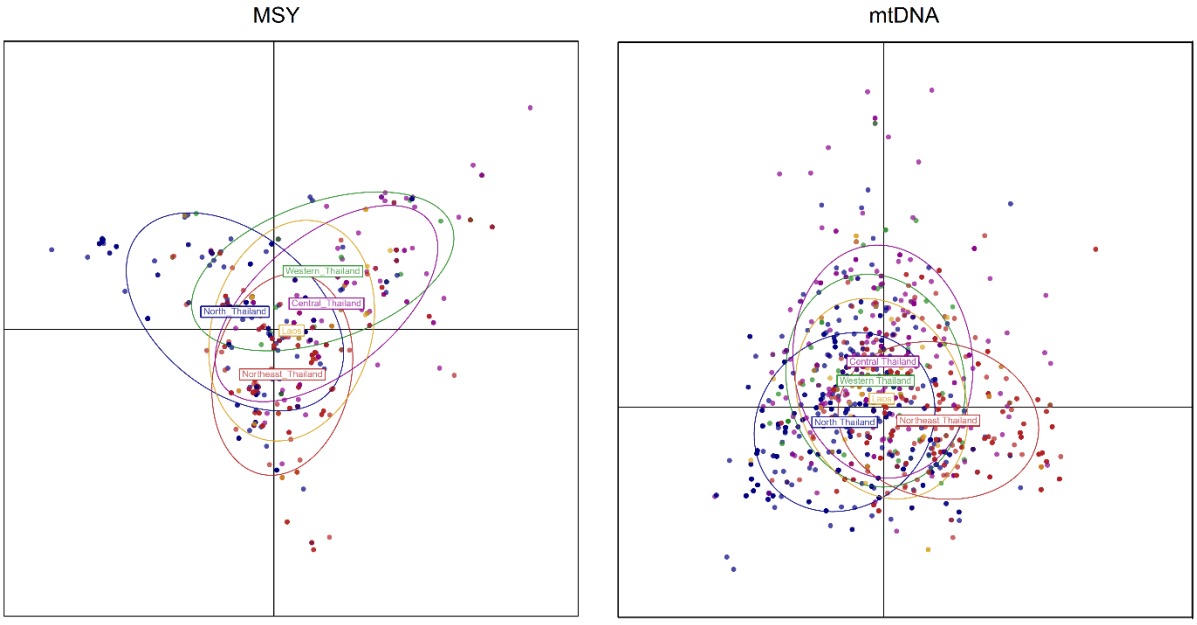
(C)



(D)



(E)



(F)

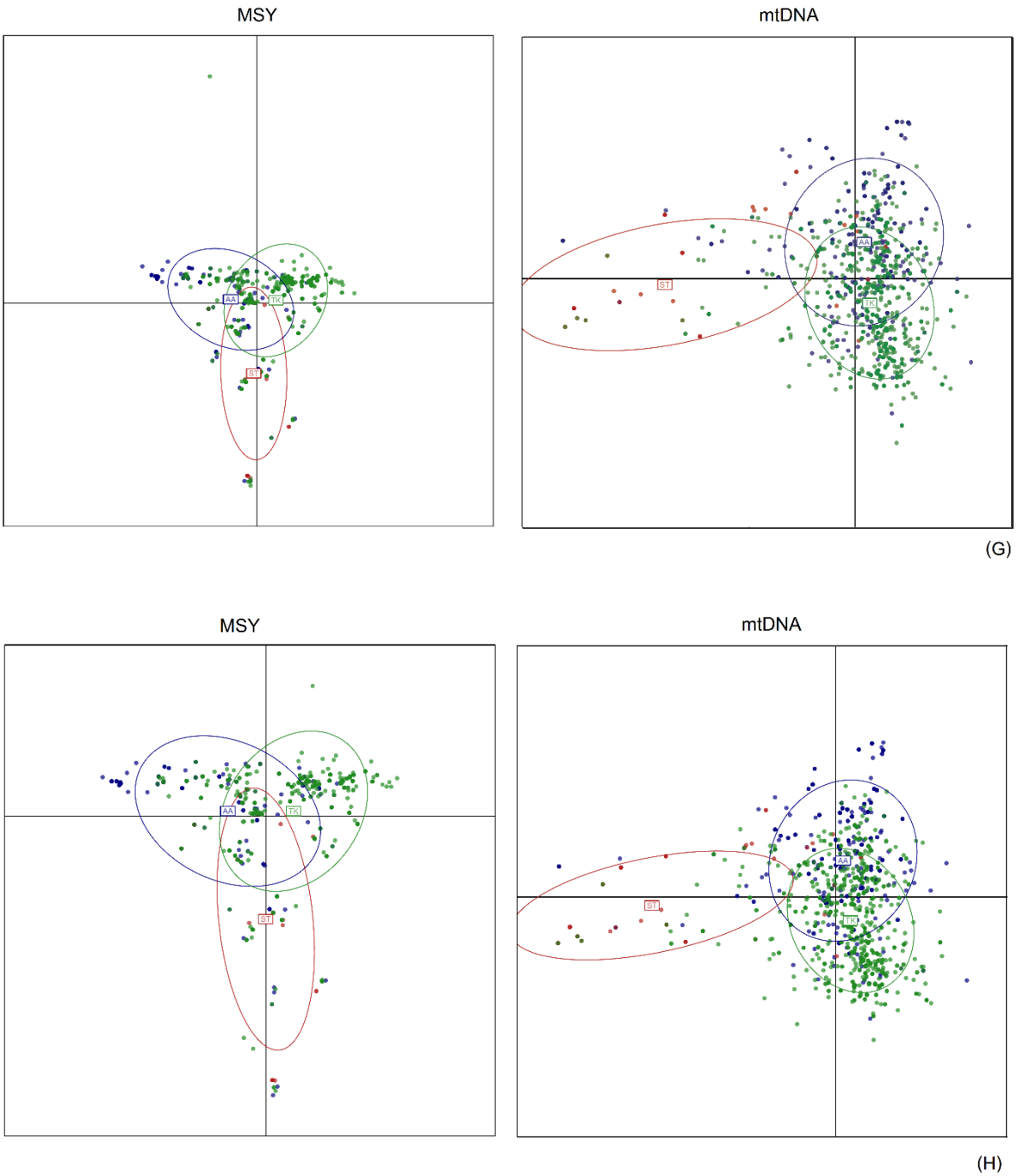
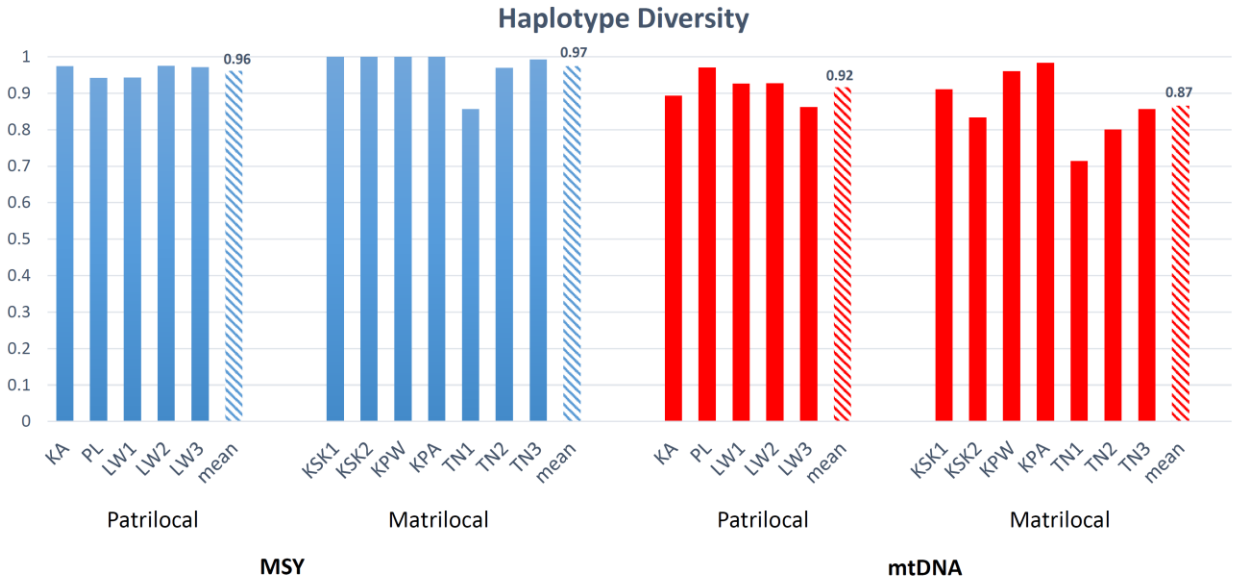
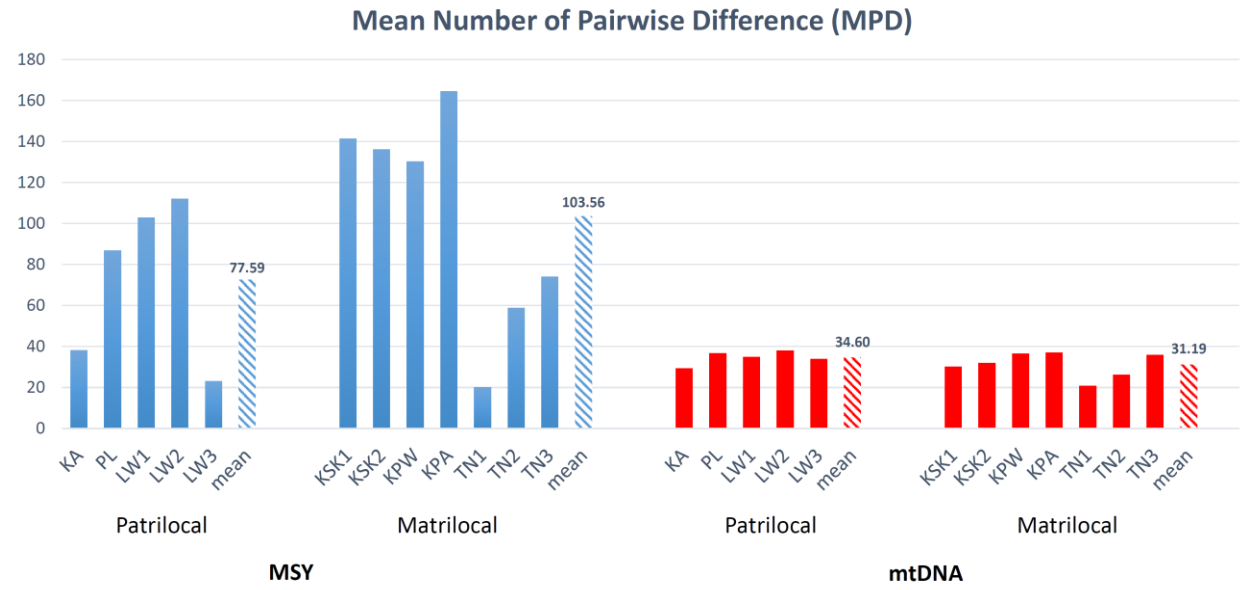


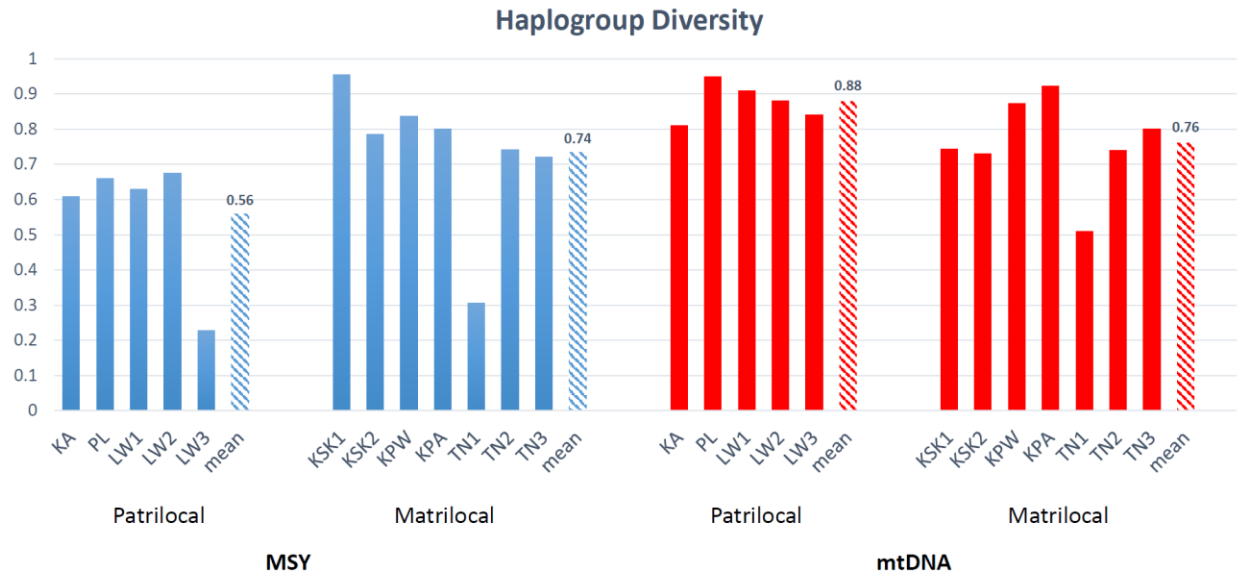
Figure S3 The DAPC results based on ethnicity, population, geography and language (A, C, E and G, respectively). The DAPC results, excluding the Maniq based on ethnicity, population, geography and language (B, D, F and H).



A

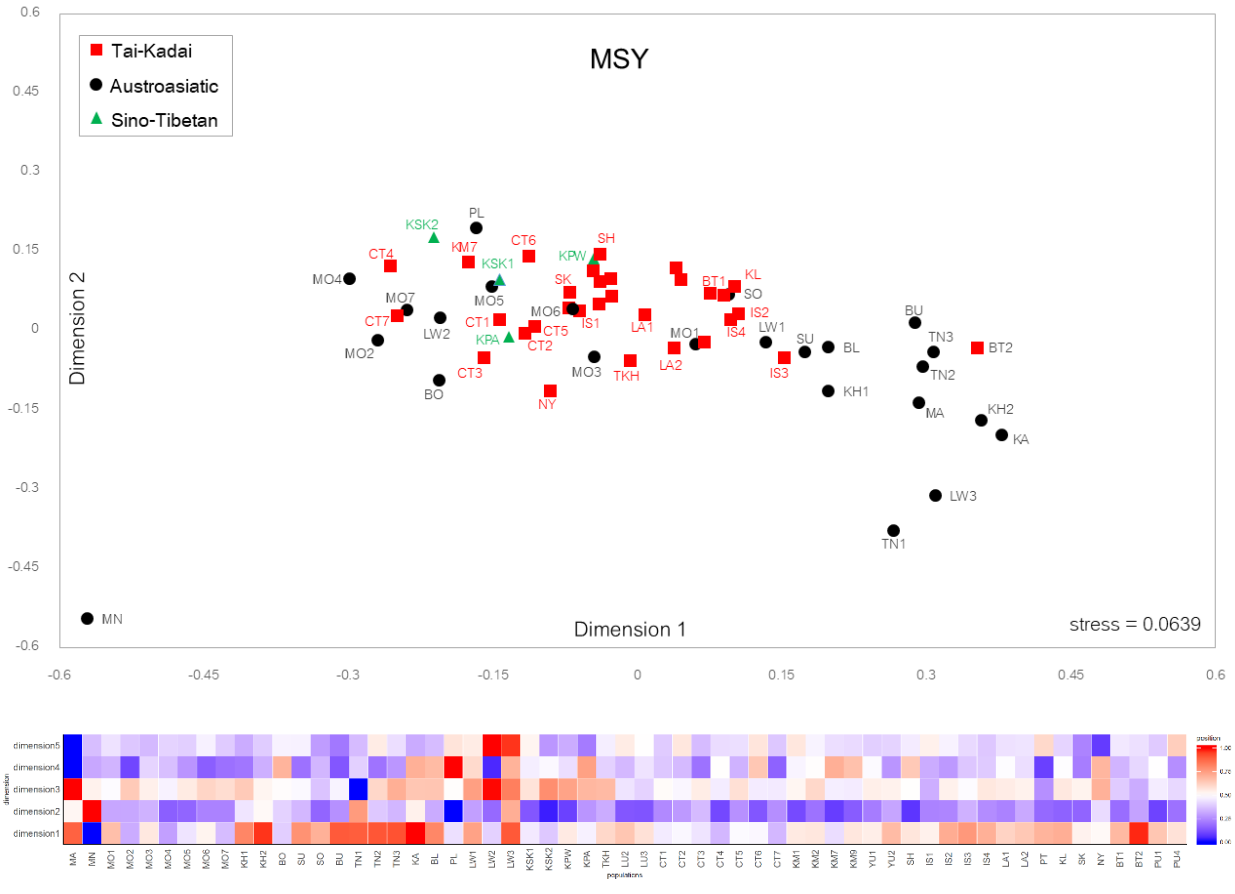


B

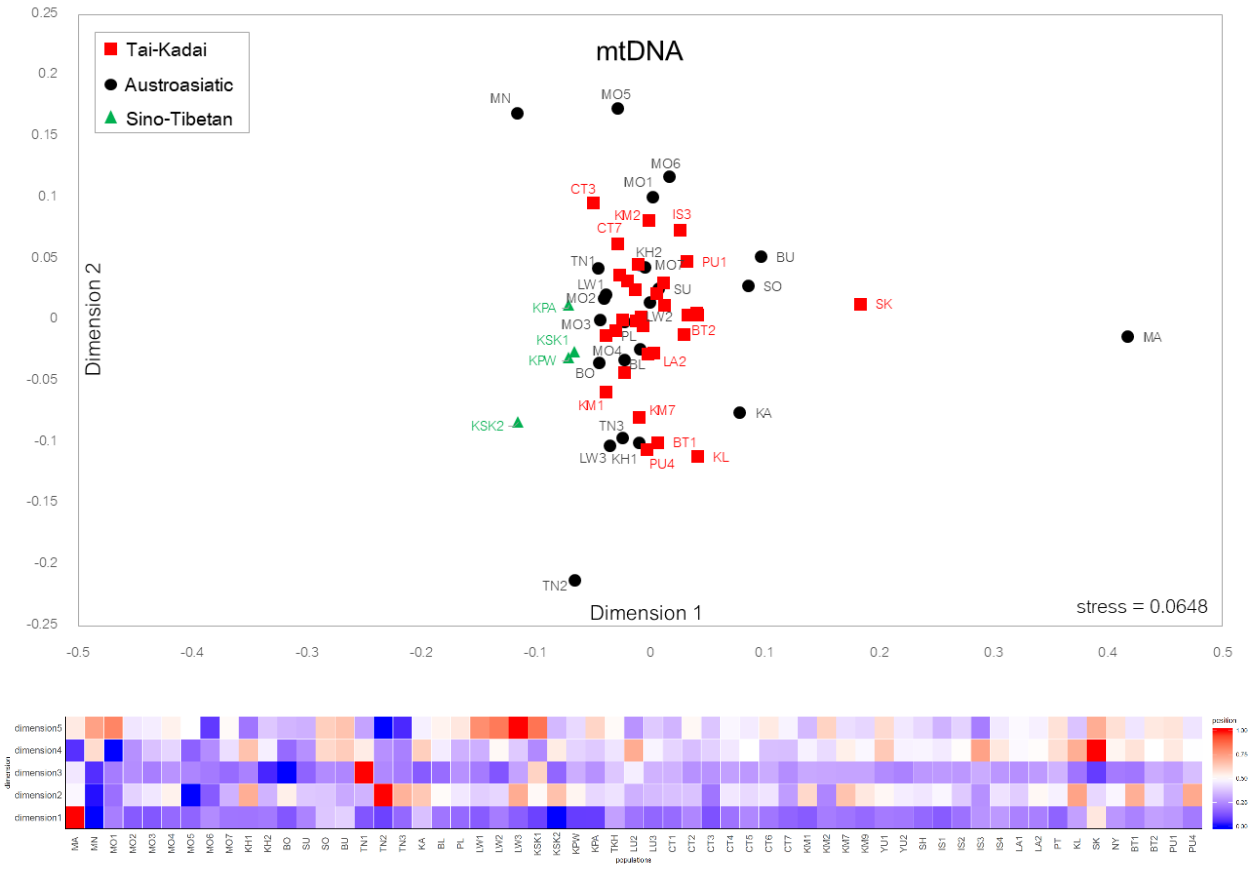


C

Figure S4 The bar plot graphs of within population genetic variation values, i.e. haplotype diversity (A), MPD (B) and haplogroup diversity (C) in patrilocal and matrilocal groups. The shaded bar in each group indicates the mean diversity in each group.

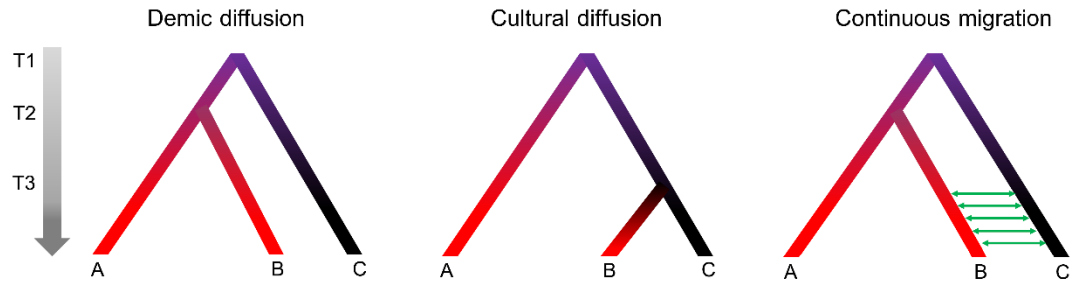


(A)



(B)

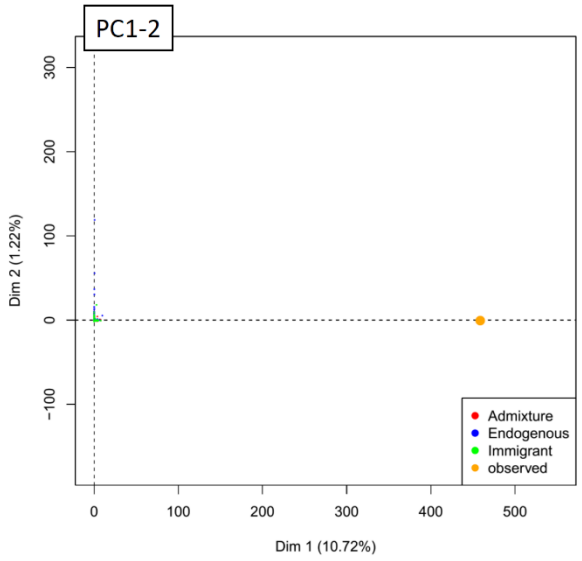
Figure S5 The MDS plot and associated heat plot based on the Φ_{st} distance matrix calculated from the dataset for 59 populations, for the MSY (A) and mtDNA (B). Population abbreviations are in Figure 1and Table S5.

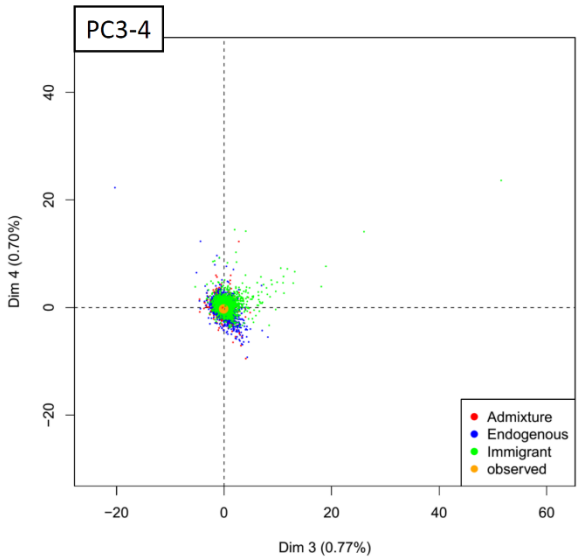


Model tests	A	B	C
Test 1	Xishuanbanna Dai	Khon Mueang	Lawa and Mon
Test 2	Xishuanbanna Dai	Lao Isan and Laotian	Suay, Khmu, Bru, Soa and Khmer
Test 3	Xishuanbanna Dai	Laotian	Suay, Khmu, Bru, Soa and Khmer
Test 4	Laotian	Lao Isan	Khmer
Test 5	Xishuanbanna Dai	Central Thais	Mon and Khmer

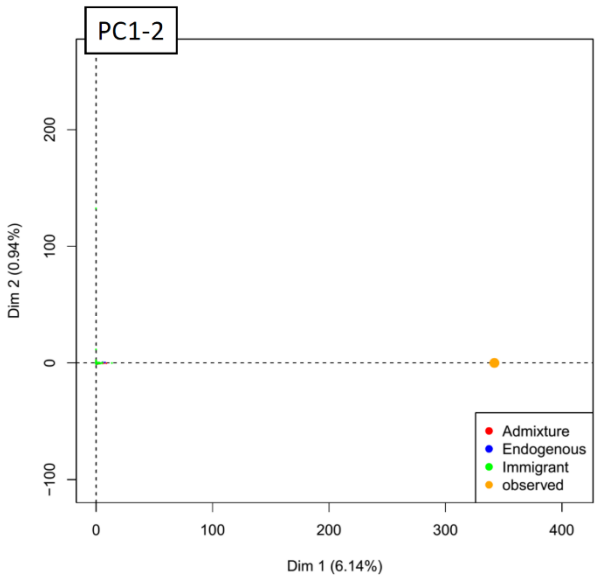
Figure S6 Three demographic models for the ABC analysis (demic diffusion, cultural diffusion and continuous migration). A, B and C represent the different populations and Test 1-5 are the different datasets used in each test. T1, T2 and T3 are either divergence time or time of gene flow.

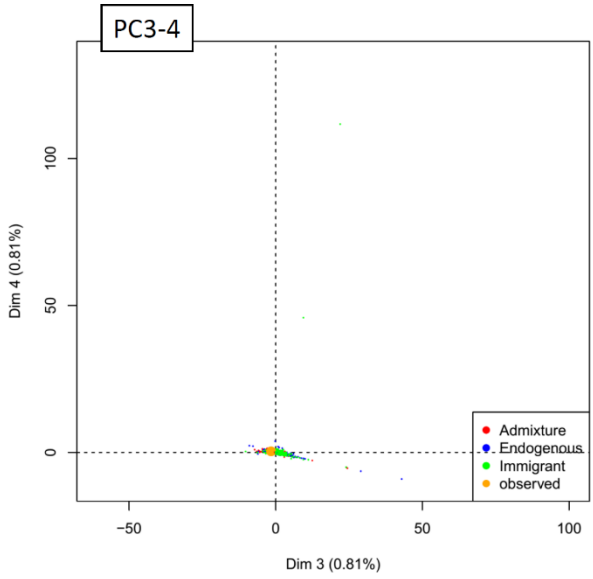
Test 1



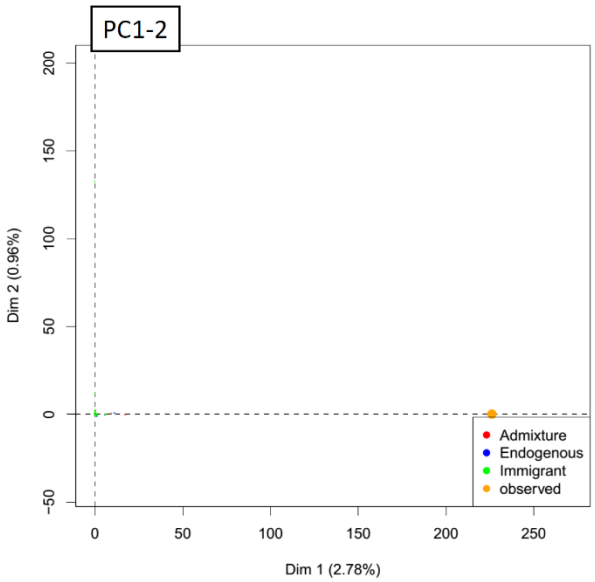


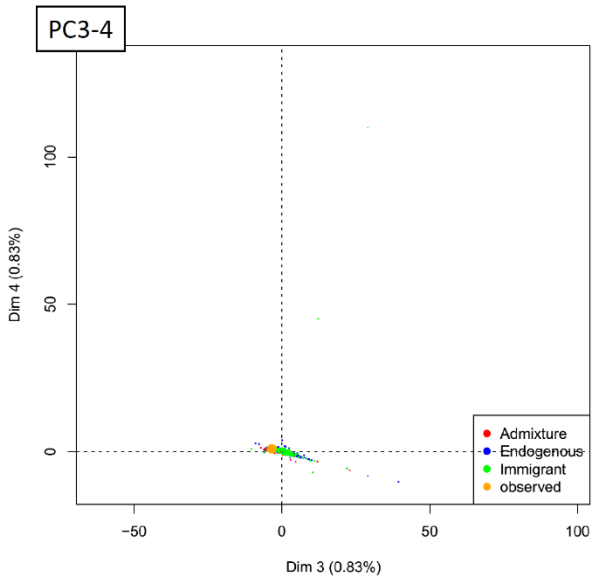
Test 2



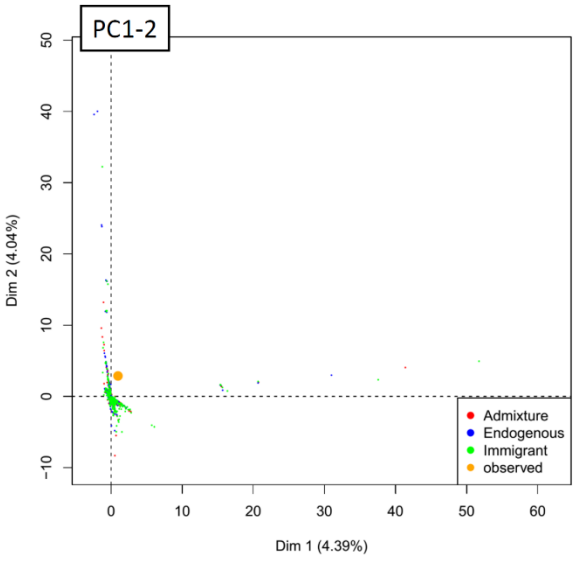


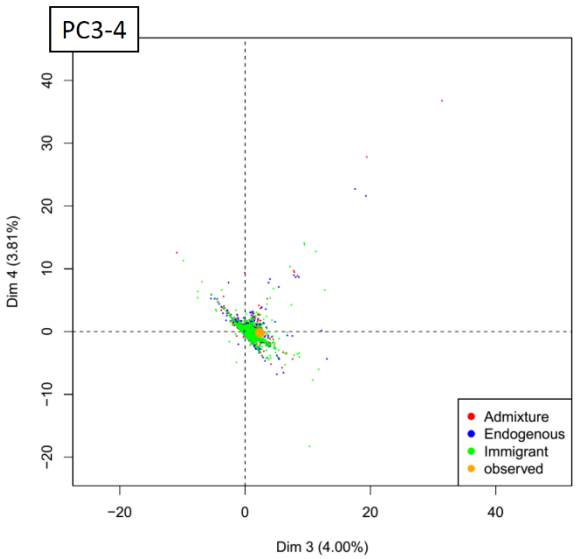
Test3





Test 4





Test 5

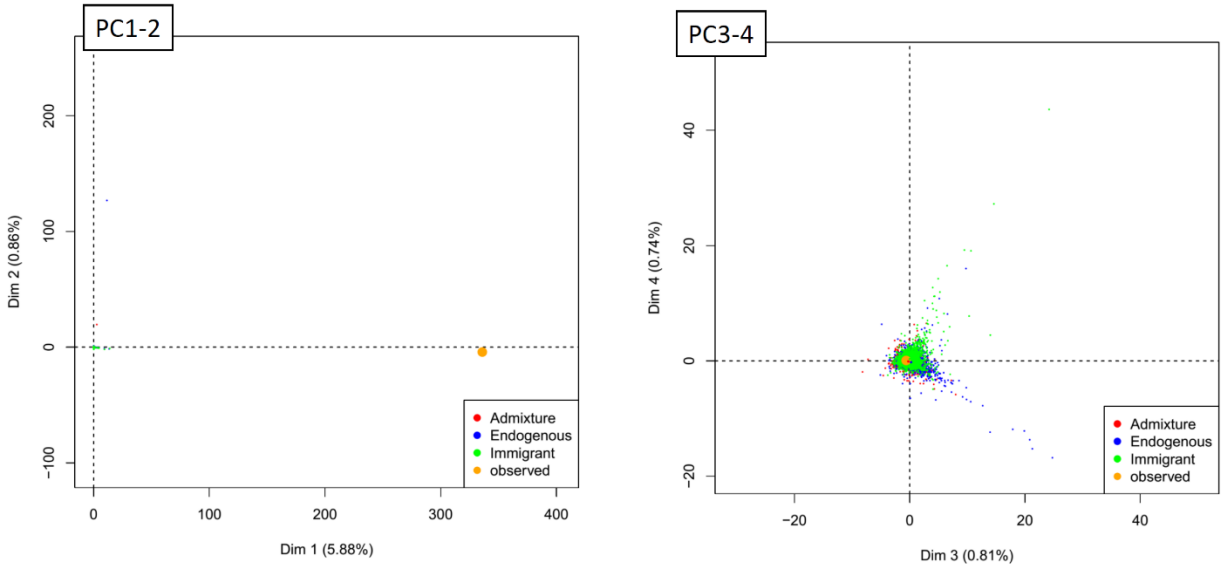
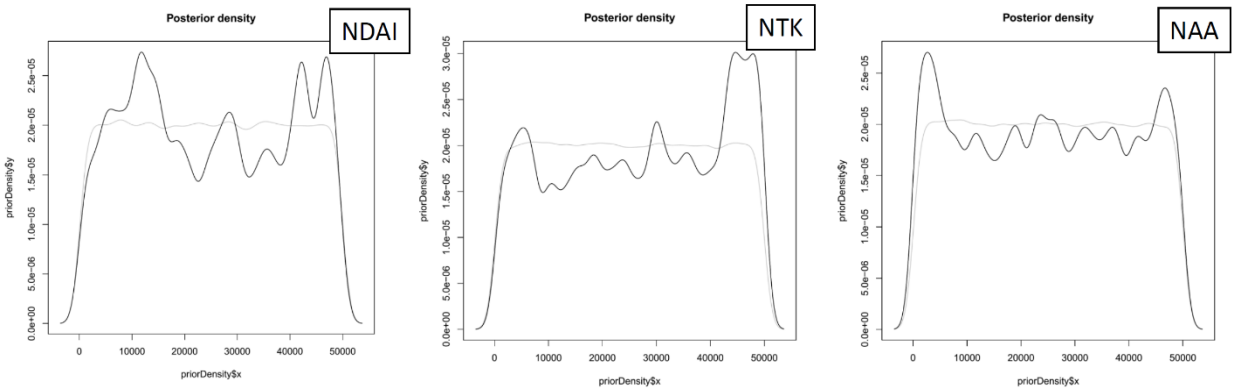
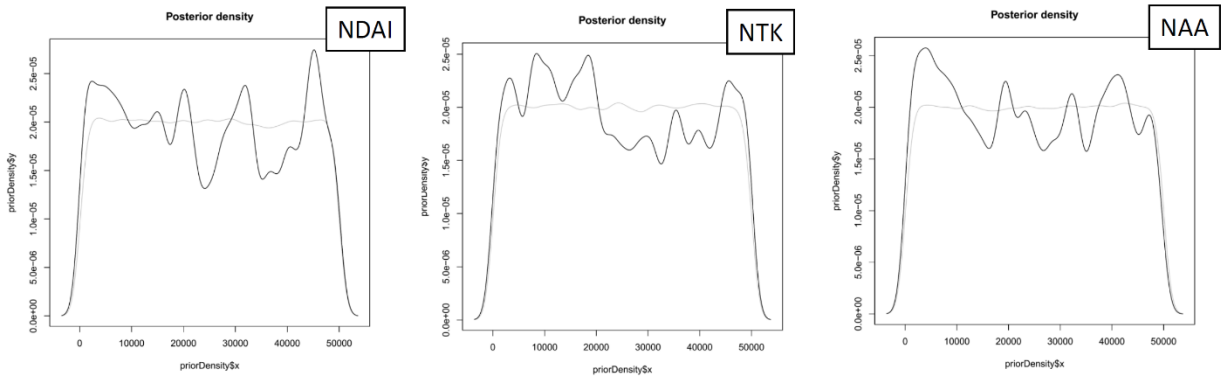


Figure S7 PCA (Principle Component Analysis) analysis based on Dimension 1 and 2 and Dimension 3 and 4 for the fit between the observed data and the simulated data generated by each model for the origin of Northern Thai (Test 1), Laotian and Lao Isan (Test 2), Laotian (Test 3), Lao Isan (Test 4) and Central Thais (Test 5).

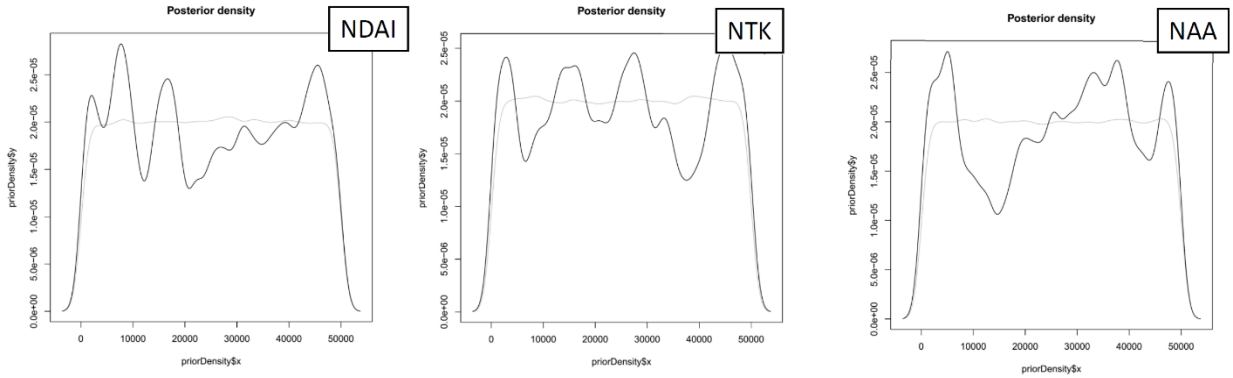
Test 1



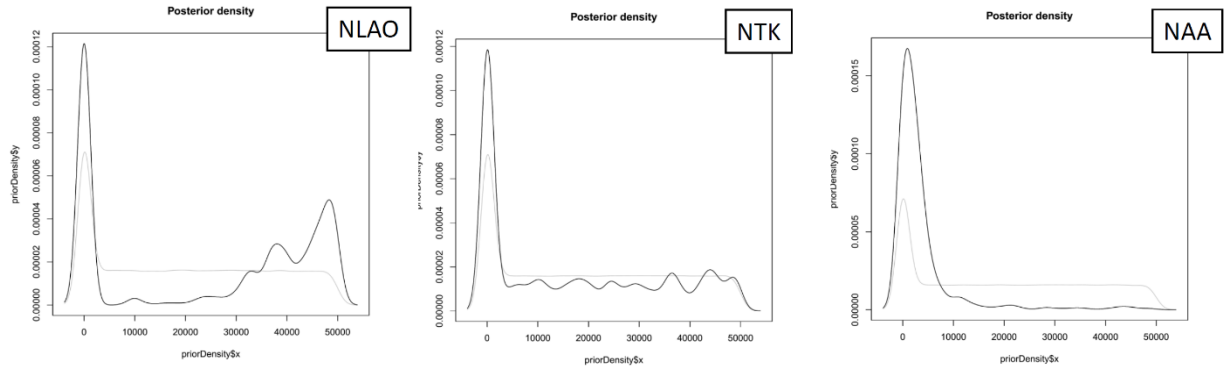
Test 2



Test 3



Test 4



Test 5

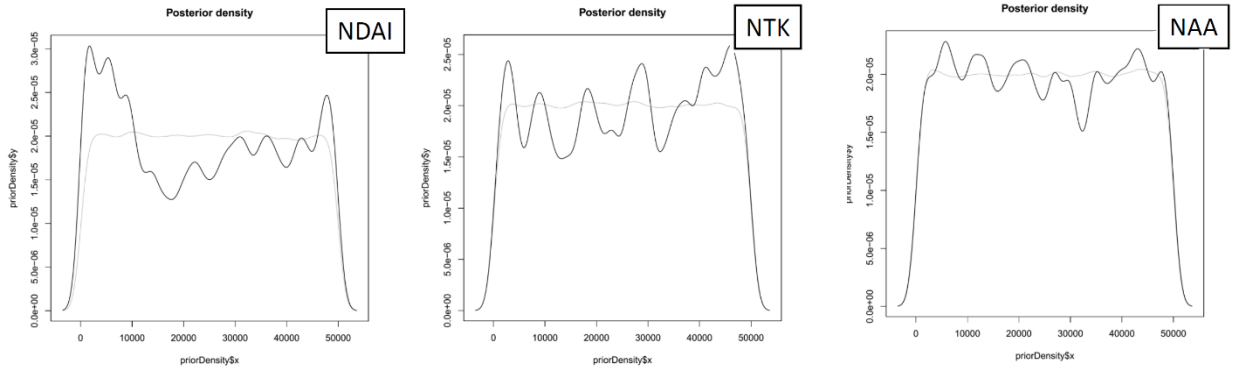


Figure S8 Graphs representing the posterior distribution of each estimated effective population sizes over the extent of the prior range (solid black) and the prior distribution of each parameter (light gray) in each model for the origin of Northern Thai (Test 1), Laotian and Lao Isan (Test 2), Laotian (Test 3), Lao Isan (Test 4) and Central Thais (Test 5).